

Methods in
Molecular Biology 1261

Springer Protocols



Raymond J. Owens *Editor*

Structural Proteomics

High-Throughput Methods

Second Edition

 Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
<http://www.springer.com/series/7651>

Structural Proteomics

High-Throughput Methods

Second Edition

Edited by

Raymond J. Owens

Oxford Protein Production Facility UK, Research Complex at Harwell, Oxfordshire, UK

 **Humana Press**

Editor

Raymond J. Owens
Oxford Protein Production Facility UK
Research Complex at Harwell, Oxfordshire, UK

ISSN 1064-3745 ISSN 1940-6029 (electronic)
ISBN 978-1-4939-2229-1 ISBN 978-1-4939-2230-7 (eBook)
DOI 10.1007/978-1-4939-2230-7
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2014956518

© Springer Science+Business Media New York 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Humana Press is a brand of Springer
Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Structural proteomics is the comprehensive analysis of protein structures in the context of whole proteomes. The field originally emerged in response to the acceleration of genomic sequencing and the decoding of the human genome and that of many human pathogens. Over the last 10 years, structural proteomics has become a mature discipline focused on specific structural biological problems or the proteomes of specific cells or environments.

One of the main deliverables of structural proteomics has been a set of technologies for the production, characterization, and structural determination of proteins with increased efficiency and at higher throughput. Many of these methods have become adopted widely in the structural biology community such as protein crystallization at the nanoliter scale. These developments have been enabled and facilitated by the increased use of automation to sustain repetitive tasks.

Since the last edition of *Methods in Molecular Biology* focused on Structural Proteomics, protocols have been improved and refined and applied to particularly challenging proteins, notably integral membrane proteins and multiprotein complexes. It is therefore very timely to take stock of current developments and collate some of the more recent methods. In Part I, the resources available for curation, annotation, and disorder prediction in silico are reviewed. Systematic recording of experimental results is critical for subsequent data mining, and the implementation of a laboratory information management system for protein production is also described in this section. Parts II and III are focused on methods for sample preparation of both proteins and crystals, which feed into structural characterization techniques reviewed in Part IV. The production of soluble proteins is considered in Part II with an emphasis on methods to express protein complexes and cell surface and secreted glycoproteins. For integral membrane proteins, the expression screening approach developed for soluble proteins has proved useful in the identification of well-expressed and stable proteins suitable for structural studies. Examples of these protocols are described in Part III. While protein crystallization remains largely an empirical process, important lessons can be learnt from past experience of both soluble and membrane proteins as reviewed in Parts II and III, respectively. In particular, paying attention to the formulation of the protein sample and choice of crystallization strategy contributes to successful crystallization experiments.

In Part IV, recent developments in high-throughput methods at synchrotrons are described as exemplified by in situ diffraction screening of crystals and CD analysis of samples in 96-well formats at the Diamond Light Source, UK. Both solution and solid-state NMR methods are also reviewed in this section. Increasingly, structural analysis of proteins by either crystallography or NMR is complemented by other techniques, and advances in mass spectrometry are described in the last chapter in Part IV.

I am grateful to all the contributors to this book for sharing their experience and expertise. I would also like to thank the *Methods in Molecular Biology* series editor, John Walker, for his guidance in preparing this volume and Springer for the opportunity of editing the second edition.

Harwell, UK

Raymond J. Owens

Contents

<i>Preface</i>	<i>v</i>
<i>Contributors</i>	<i>ix</i>
PART I BIOINFORMATICS AND DATA MANAGEMENT	
1 Protein Structure Annotation Resources <i>Margaret J. Gabanyi and Helen M. Berman</i>	3
2 PiMS: A Data Management System for Structural Proteomics <i>Chris Morris</i>	21
3 Prediction and Analysis of Intrinsically Disordered Proteins <i>Marco Punta, István Simon, and Zsuzsanna Dosztányi</i>	35
PART II SOLUBLE PROTEINS	
4 Characterization and Production of Protein Complexes by Co-expression in <i>Escherichia coli</i> <i>Matthias Haffke, Martin Marek, Martin Pelosse, Marie-Laure Diebold, Uwe Schlattner, Imre Berger, and Christophe Romier</i>	63
5 The Production of Multiprotein Complexes in Insect Cells Using the Baculovirus Expression System <i>Wassim Abdulrahman, Laura Radu, Frederic Garzoni, Olga Kolesnikova, Kapil Gupta, Judit Osz-Papai, Imre Berger, and Arnaud Poterszman</i>	91
6 Production of Cell Surface and Secreted Glycoproteins in Mammalian Cells <i>Elena Seiradake, Yuguang Zhao, Weixian Lu, A. Radu Aricescu, and E. Yvonne Jones</i>	115
7 Cell-Free Protein Synthesis Systems Derived from Cultured Mammalian Cells <i>Andreas K. Brödel, Doreen A. Wüstenhagen, and Stefan Kubick</i>	129
8 Crystallization: Digging into the Past to Learn Lessons for the Future. <i>Vincent J. Fazio, Thomas S. Peat, and Janet Newman</i>	141
PART III MEMBRANE PROTEINS	
9 Screening of Stable G-Protein-Coupled Receptor Variants in <i>Saccharomyces cerevisiae</i> <i>Mitsunori Shiroishi and Takuya Kobayashi</i>	159
10 Cell-Free Expression of G-Protein-Coupled Receptors <i>Erika Orbán, Davide Proverbio, Stefan Haberstock, Volker Dötsch, and Frank Bernhard</i>	171

11	GFP-Based Expression Screening of Membrane Proteins in Insect Cells Using the Baculovirus System	197
	<i>Nien-Jen Hu, Heather Rada, Nahid Rahman, Joanne E. Nettleship, Louise Bird, So Iwata, David Drew, Alexander D. Cameron, and Raymond J. Owens</i>	
12	Methods for the Successful Crystallization of Membrane Proteins	211
	<i>Isabel Moraes and Margarida Archer</i>	
PART IV STRUCTURAL CHARACTERIZATION OF PROTEINS		
13	Application of In Situ Diffraction in High-Throughput Structure Determination Platforms.	233
	<i>Pierre Aller, Juan Sanchez-Weatherby, James Foadi, Graeme Winter, Carina M.C. Lobley, Danny Axford, Alun W. Ashton, Domenico Bellini, Jose Brandao-Neto, Simone Culurgioni, Alice Douangamath, Ramona Duman, Gwyndaf Evans, Stuart Fisher, Ralf Flaig, David R. Hall, Petra Lukacik, Marco Mazzorana, Katherine E. McAuley, Vitaliy Mykhaylyk, Robin L. Owen, Neil G. Paterson, Pierpaolo Romano, James Sandy, Thomas Sorensen, Frank von Delft, Armin Wagner, Anna Warren, Mark Williams, David I. Stuart, and Martin A. Walsh</i>	
14	CD Spectroscopy: An Essential Tool for Quality Control of Protein Folding.	255
	<i>Giuliano Siligardi and Rohanah Hussain</i>	
15	High-Throughput Studies of Protein Shapes and Interactions by Synchrotron Small-Angle X-Ray Scattering.	277
	<i>Cy M. Jeffries and Dmitri I. Svergun</i>	
16	Automated Structure Determination from NMR Spectra.	303
	<i>Elena Schmidt and Peter Güntert</i>	
17	Solid-State Nuclear Magnetic Resonance Spectroscopy for Membrane Protein Structure Determination	331
	<i>Peter J. Judge, Garrick F. Taylor, Hugh R.W. Dannatt, and Anthony Watts</i>	
18	Native Mass Spectrometry: Towards High-Throughput Structural Proteomics	349
	<i>Frances D.L. Kondrat, Weston B. Struwe, and Justin L.P. Benesch</i>	
	<i>Index</i>	373

Contributors

- WASSIM ABDULRAHMAN • *Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS/INSERM/UDS, Illkirch, France*
- PIERRE ALLER • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- MARGARIDA ARCHER • *Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa (ITQB-UNL), Oeiras, Portugal*
- A. RADU ARICESCU • *The Division of Structural Biology, The Henry Wellcome Building for Genomic Medicine, Oxford, UK*
- ALUN W. ASHTON • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- DANNY AXFORD • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- DOMENICO BELLINI • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- JUSTIN L.P. BENESCH • *Department of Chemistry, Physical and Theoretical Chemistry Laboratory, University of Oxford, Oxford, UK*
- IMRE BERGER • *Grenoble Outstation and Unit of Virus Host-Cell Interactions (UVHCI), European Molecular Biology Laboratory (EMBL), Université Grenoble Alpes, EMBL-CNRS, Grenoble, France*
- HELEN M. BERMAN • *Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey, Piscataway, NJ, USA*
- FRANK BERNHARD • *Centre for Biomolecular Magnetic Resonance, Institute of Biophysical Chemistry, Goethe-University of Frankfurt/Main, Frankfurt/Main, Germany*
- LOUISE BIRD • *Oxford Protein Production Facility-UK, Research Complex at Harwell, R92 Rutherford Appleton Laboratory, Harwell, UK*
- JOSE BRANDAO-NETO • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- ANDREAS K. BRÖDEL • *Fraunhofer Institute for Biomedical Engineering (IBMT), Potsdam, Germany*
- ALEXANDER D. CAMERON • *School of Life Sciences, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK*
- SIMONE CULURGIONI • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- HUGH R.W. DANNATT • *Biomembrane Structure Unit, Department of Biochemistry, University of Oxford, Oxford, UK*
- FRANK VON DELFT • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- MARIE-LAURE DIEBOLD • *Département de Biologie Structurale Intégrative, Centre de Biologie Intégrative, Institut de Génétique et Biologie Moléculaire et Cellulaire (IGBMC), UDS, CNRS, INSERM, Illkirch, France*
- ZSUZSANNA DOSZTÁNYI • *Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary*
- VOLKER DÖTSCH • *Centre for Biomolecular Magnetic Resonance, Institute of Biophysical Chemistry, Goethe-University of Frankfurt/Main, Frankfurt/Main, Germany*
- ALICE DOUANGAMATH • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- DAVID DREW • *Centre for Biomembrane Research, Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden*
- RAMONA DUMAN • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*

- GWYNDAF EVANS • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- VINCENT J. FAZIO • *Collaborative Crystallisation Centre, CSIRO Materials Science and Engineering, Parkville, VIC, Australia*
- STUART FISHER • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- RALF FLAIG • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- JAMES FOADI • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- MARGARET J. GABANYI • *Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey, Piscataway, NJ, USA*
- FREDERIC GARZONI • *Grenoble Outstation and Unit of Virus Host-Cell Interactions (UVHCI), European Molecular Biology Laboratory (EMBL), Universite Grenoble Alpes, EMBL-CNRS, Grenoble, France*
- PETER GÜNTERT • *Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, and Frankfurt Institute of Advanced Studies, Frankfurt am Main, Germany; Graduate School of Science and Engineering, Tokyo Metropolitan University, Hachioji, Tokyo, Japan*
- KAPIL GUPTA • *Grenoble Outstation and Unit of Virus Host-Cell Interactions (UVHCI), European Molecular Biology Laboratory (EMBL), Universite Grenoble Alpes, EMBL-CNRS, Grenoble, France*
- STEFAN HABERSTOCK • *Centre for Biomolecular Magnetic Resonance, Institute of Biophysical Chemistry, Goethe-University of Frankfurt/Main, Frankfurt/Main, Germany*
- MATTHIAS HAEFFKE • *Grenoble Outstation and Unit of Virus Host-Cell Interactions (UVHCI), European Molecular Biology Laboratory (EMBL), Universite Grenoble Alpes, EMBL-CNRS, Grenoble, France*
- DAVID R. HALL • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- NIEN-JEN HU • *Institute of Biochemistry, National Chung Hsing University, Taichung, Taiwan*
- ROHANAH HUSSAIN • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- SO IWATA • *Division of Molecular Biosciences, Imperial College London, London, UK*
- CY M. JEFFRIES • *Hamburg Outstation c/o DESY, European Molecular Biology Laboratory (EMBL), Hamburg, Germany*
- E. YVONNE JONES • *The Division of Structural Biology, The Henry Wellcome Building for Genomic Medicine, Oxford, UK*
- PETER J. JUDGE • *Biomembrane Structure Unit, Department of Biochemistry, University of Oxford, Oxford, UK*
- TAKUYA KOBAYASHI • *Department of Cell Biology, Graduate School of Medicine, Kyoto University, Kyoto, Japan*
- OLGA KOLESNIKOVA • *Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS/INSERM/UDS, Illkirch, France*
- FRANCES D.L. KONDRAT • *Department of Chemistry, Physical and Theoretical Chemistry Laboratory, University of Oxford, Oxford, UK*
- STEFAN KUBICK • *Fraunhofer Institute for Biomedical Engineering (IBMT), Potsdam, Germany*
- CARINA M.C. LOBLEY • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- WEIXIAN LU • *The Division of Structural Biology, The Henry Wellcome Building for Genomic Medicine, Oxford, UK*
- PETRA LUKACIK • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- MARTIN MAREK • *Département de Biologie Structurale Intégrative, Centre de Biologie Intégrative, Institut de Génétique et Biologie Moléculaire et Cellulaire (IGBMC), UDS, CNRS, INSERM, Illkirch, France*
- MARCO MAZZORANA • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*

- KATHERINE E. MCAULEY • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- ISABEL MORAES • *Department of Life Sciences, Imperial College London, London, UK;
Membrane Protein Laboratory, Diamond Light Source, Didcot, Chilton, UK*
- CHRIS MORRIS • *STFC, Daresbury Laboratory, Sci-Tech Daresbury, Daresbury, Warrington, UK*
- VITALIY MYKHAYLYK • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- JOANNE E. NETTLESHIP • *Oxford Protein Production Facility-UK, Research Complex at
Harwell, R92 Rutherford Appleton Laboratory, Harwell, UK*
- JANET NEWMAN • *Collaborative Crystallisation Centre, CSIRO Biomedical
Manufacturing, Parkville, VIC, Australia*
- ERIKA ORBÁN • *Centre for Biomolecular Magnetic Resonance, Institute of Biophysical
Chemistry, Goethe-University of Frankfurt/Main, Frankfurt/Main, Germany*
- JUDIT OSZ-PAPAI • *Institut de Génétique et de Biologie Moléculaire et Cellulaire,
CNRS/INSERM/UDS, Illkirch, France*
- ROBIN L. OWEN • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- RAYMOND J. OWENS • *Oxford Protein Production Facility-UK, Research Complex at
Harwell, R92 Rutherford Appleton Laboratory, Harwell, UK*
- NEIL G. PATERSON • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- THOMAS S. PEAT • *Collaborative Crystallisation Centre, CSIRO Materials Science
and Engineering, Parkville, VIC, Australia*
- MARTIN PELOSSE • *Grenoble Outstation and Unit of Virus Host-Cell Interactions
(UVHCI), European Molecular Biology Laboratory (EMBL), Université Grenoble Alpes,
EMBL-CNRS, Grenoble, France*
- JONATHAN POISSON • *Department of Computational Medicine and Bioinformatics,
University of Michigan, Ann Arbor, MI, USA*
- ARNAUD POTERSZMAN • *Institut de Génétique et de Biologie Moléculaire et Cellulaire,
CNRS/INSERM/UDS, Illkirch, France*
- DAVIDE PROVERBIO • *Centre for Biomolecular Magnetic Resonance, Institute of Biophysical
Chemistry, Goethe-University of Frankfurt/Main, Frankfurt/Main, Germany*
- MARCO PUNTA • *European Molecular Biology Laboratory, European Bioinformatics
Institute (EMBL-EBI), Hinxton, Cambridge, UK; Wellcome Trust Sanger Institute,
Hinxton, Cambridge, UK*
- HEATHER RADA • *Oxford Protein Production Facility-UK, Research Complex at Harwell,
R92 Rutherford Appleton Laboratory, Harwell, UK*
- LAURA RADU • *Institut de Génétique et de Biologie Moléculaire et Cellulaire,
CNRS/INSERM/UDS, Illkirch, France*
- NAHID RAHMAN • *Oxford Protein Production Facility-UK, Research Complex at Harwell,
R92 Rutherford Appleton Laboratory, Harwell, UK*
- PIERPAOLO ROMANO • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- CHRISTOPHE ROMIER • *Département de Biologie Structurale Intégrative, Centre de Biologie
Intégrative, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC),
UDS, CNRS, INSERM, Illkirch, France*
- AMBRISH ROY • *Department of Computational Medicine and Bioinformatics, University
of Michigan, Ann Arbor, MI, USA*
- JUAN SANCHEZ-WEATHERBY • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- JAMES SANDY • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- UWE SCHLATTNER • *Laboratory of Fundamental and Applied Bioenergetics (LBFA),
Université Grenoble Alpes, INSERM, Grenoble, France*

- ELENA SCHMIDT • *Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, and Frankfurt Institute of Advanced Studies, Frankfurt am Main, Germany*
- ELENA SEIRADAKE • *The Division of Structural Biology, The Henry Wellcome Building for Genomic Medicine, Oxford, UK*
- MITSUNORI SHIROISHI • *Graduate School of Pharmaceutical Sciences, Kyushu University, Fukuoka, Japan*
- GIULIANO SILIGARDI • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- ISTVÁN SIMON • *Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary*
- THOMAS SORENSEN • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- WESTON B. STRUWE • *Department of Chemistry, Physical and Theoretical Chemistry Laboratory, University of Oxford, Oxford, UK*
- DAVID I. STUART • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- DMITRI I. SVERGUN • *Hamburg Outstation c/o DESY, European Molecular Biology Laboratory (EMBL), Hamburg, Germany*
- GARRICK F. TAYLOR • *Biomembrane Structure Unit, Department of Biochemistry, University of Oxford, Oxford, UK*
- ARMIN WAGNER • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- MARTIN A. WALSH • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- ANNA WARREN • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- ANTHONY WATTS • *Biomembrane Structure Unit, Department of Biochemistry, University of Oxford, Oxford, UK*
- MARK WILLIAMS • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- GRAEME WINTER • *Diamond Light Source Ltd, Chilton, Oxfordshire, UK*
- DOREEN A. WÜSTENHAGEN • *Fraunhofer Institute for Biomedical Engineering (IBMT), Potsdam, Germany*
- DONG XU • *Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA*
- RENXIANG YAN • *Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA*
- JIANXI YANG • *Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA*
- YANG ZHANG • *Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA*
- YUGUANG ZHAO • *The Division of Structural Biology, The Henry Wellcome Building for Genomic Medicine, Oxford, UK*

Part I

Bioinformatics and Data Management

Chapter 1

Protein Structure Annotation Resources

Margaret J. Gabanyi and Helen M. Berman

Abstract

A key reason three-dimensional (3-D) protein structures are annotated with supporting or derived information is to understand the molecular basis of protein function. To this end, protein structure annotation databases curate key facts and observations, based on community-accepted standards, about the ~100,000 3-D experimental protein structures to date. This review will introduce the primary structure repositories, databases, and value-added structural annotation databases, as well as the range of information they provide. The different levels of annotation data (primary vs. derived vs. inferred) and how they should all be considered accordingly will also be described.

Key words Crystallography, Databases, 3-D electron microscopy, Nuclear magnetic resonance, Small angle scattering, Structural biology, Structural proteomics

1 Introduction

There is a vast amount of biological data available in public repositories. At the time this review was prepared, there were over 157 billion genes registered in GenBank (v. 204, Oct 2014) [1], half a million Swiss-Prot (curated) protein sequences and 87 million TrEMBL (computationally annotated) protein sequences in UniProtKB (v. 2014_10) [2], and over 100,000 three-dimensional macromolecular structures in the Protein Data Bank archive (v. 2014-10-28) [3, 4]. Improved high-throughput technologies have helped gather new sequence information faster than we are able to analyze it. There is simply not enough time, funding, or skilled workforce to study every new protein in the same detail. Therefore, to maximize the value of the information garnered from the genes, proteins, and structures that we do study, annotations are captured so that they may be applied to other similar (*homologous*) proteins.

Since the 3-D structure of proteins are ultimately responsible for most actions in the cell, this review will focus on protein structure annotation resources that aim to explain a molecular

basis of protein function. First, we will describe the Protein Data Bank archive that houses the primary three-dimensional structural data and annotations for proteins and nucleic acids. We will then review the principal databases and resources that allow access to the PDB data for visualization and analysis, as well as several new repositories addressing new methods in structural biology. Next, structural classification and annotation databases that aim to categorize observations made by the structural biology community will be described. Selected specialty databases that focus on a specific biological aspect will also be covered. Lastly, we will review data aggregators and community-driven structure annotation projects that work to integrate protein structures with functional annotation to enable a better understanding of living systems and disease.

There is a hierarchy to the types of annotations that are captured and how reliable they are. At the top level is the structural data itself, composed usually of coordinates or 3-D maps. Next are the primary annotations, which have either been experimentally measured and verified or report experimental parameters, so that the study is reproducible. Next follow the derived annotations, which are calculated from experimental measurements. Last come the annotations that are propagated from homologous protein sequences or structures. It is important for biologists to understand the strengths and caveats of these data so that they can be applied to their maximum benefit. For this reason, it is important that the community experts be involved with the determination of these annotations to ensure consistency and reliability of the data for everyone.

2 The Protein Data Bank (PDB) Archive

The *PDB* is the primary worldwide data archive for experimentally determined structures of biological macromolecules. It was established in 1971 following several years of discussion by the community who felt that such data should be openly available to the entire biological research community [5]. At the 1971 Cold Spring Harbor Symposium *Structure and Function of Proteins at the Three-Dimensional Level*, Walter Hamilton of Brookhaven National Laboratory agreed to start the archive in collaboration with Olga Kennard of the Cambridge Structural Database; at that time, there were less than a dozen protein structures. Now 40+ years later, there are more than 100,000 structures in the PDB, and coordinates files are downloaded more than one million times every day.

Since the mid-2000s, the PDB archive has been managed by the worldwide Protein Data Bank (wwPDB) consortium [4], composed of four international members dedicated to maintaining a high-quality, uniformly curated resource: The Research

Collaboratory for Structural Bioinformatics PDB (RCSB PDB, USA) [3], PDB in Europe (PDBe, UK) [6], PDB in Japan (PDBj, Japan) [7], and the BioMagResBank (BMRB, USA) [8]. Being a community resource, the wwPDB regularly establishes task forces and meets with field experts who make recommendations to address growing complexities in the data. Recent developments made in response to these expert recommendations include the ongoing implementation of latest validation standards into data validation and processing procedures [9–11], the availability of structure validation reports for journal referees during peer-review [12], and, very recently, the shift toward the mmCIF/PDBx file format from the limiting PDB format [13].

The wwPDB collects structural data and also preliminary information about the macromolecules present in the structure [14], which are checked and represented using a data dictionary [15]. For structures determined by X-ray crystallography, the wwPDB collects the x , y , z (Cartesian) coordinates for every resolved atom in a biomolecule and the original structure factor data to generate the electron density maps. For structures determined by nuclear magnetic resonance (NMR), the Cartesian coordinates for each model in the deposited ensemble and restraint and chemical shift data are collected. In addition to the coordinates and data sets, methods-based annotations that represent how the sample was made and stored such as source and host organisms and vectors, crystallization conditions, or NMR solutions are also included. Additional author-provided or calculated (derived) annotations include secondary structure element identification and symmetry and biological assembly information. Biological annotations such as GenBank and UniProt identifiers and the broad function of the macromolecule or enzyme classification are added to the record as well. Discrepancies between reported and reference sequences (such as mutations) are noted.

The next step in the annotation process is to validate the deposited atomic data. Covalent bond distances and angles, stereochemical validation, close contacts, ligand and atom nomenclature, sequence comparison, distant waters, and overall geometry are reviewed and compared by wwPDB annotators to standards recommended by the methods validation task force, and authors are informed of any inconsistencies. For X-ray structures, the fit of the refined model is also checked against its experimental data. Outliers may be noted in the entry. After data are fully reviewed and annotated, they are approved by the depositing authors for public release. The PDB archive File Transfer Protocol (FTP) server is updated each week. The wwPDB regularly reviews the archive and makes corrections to data files to improve the quality and consistency of the data across the archive. These efforts also result in improved data processing procedures and in some cases, the creation of external reference files and/or new data dictionaries.

Recent work includes improving how peptide-like inhibitors and antibiotics are represented in the PDB archive [16]. These are the first installment in a new Biologically Interesting molecule Reference Dictionary (BIRD), which contains chemical descriptions, sequence and linkage information, and functional and classification information as captured from the structures and other external sources.

It should be noted, however, that inclusion of other chemical components in the entry do not infer any function. It is usually difficult to determine if the chemical components (such as those coming from a crystallization solution) are an artifact of the experiment or they mimic some functional cofactor or substrate. The same is true for “site” records; authors can provide annotations regarding the functional role of an amino acid, but wwPDB also automatically calculates which amino acids are within a certain distance of nearby ligands and reports all of them.

A new unified Common Deposition and Annotation system has been developed by the wwPDB and has been released to the community in 2014 (<http://deposit.wwpdb.org/>) [17]. To address data quality issues, new validation metrics report how well each structure compares to those solved by the same method, and also how the structure compares to those determined to a similar resolution [9]. Validation reports for all X-ray crystallographic entries are now publicly available from the PDB FTP archive as of March 2014 (ftp://ftp.wwpdb.org/pub/pdb/validation_reports/).

3 Macromolecular Structure Resources

There are many data resources that depend wholly, or in part, on the contents of the PDB. This section will describe how each resource site utilizes the structural annotation information for search and analysis. The URL addresses for these resources are listed in Table 1.

The *RCSB PDB* provides structure search, visualization, and analysis tools to retrieve relevant entries from the PDB archive [18]. Simple searches including by PDB ID, molecule name, author name, or plain text are available. As words are entered in the search box, the autocomplete function will find individual structure entries or suggest groups of structures that share a particular attribute or annotation (molecule type, structural fold, protein family, etc.). Users can also enter a sequence or draw a ligand structure, or use the “advanced” and “browse” search features to filter the archive by combining ~80 categories of structural (i.e., folds), biological (i.e., Gene Ontology [GO]), or experimental (i.e., method) annotations. Unless looking for a specific entry, the result is a list of relevant structures that can be organized into customizable reports for future reference.

Table 1
List of structural resources selected for this review

Repository for 3-D biological macromolecule structures Protein Data Bank [3]	Macromolecule structure coordinates archive	http://www.wwpdb.org
Structure resources		
RCSB Protein Data Bank [3]	Macromolecular structure deposition, search, and analysis tools (USA)	http://www.rcsb.org
PDB in Europe (PDBe) [6]	Macromolecular structure deposition, search, and analysis tools (Europe)	http://www.pdbe.org
PDB in Japan (PDBj) [7]	Macromolecular structure deposition, search, and analysis tools (Japan)	http://www.pdbj.org
BioMagResBank (BMRB) [8]	NMR data set deposition, search, and validation tools	http://www.bmrb.wisc.edu
EMDataBank [29]	Cryo-EM data deposition, search, and analysis tools	http://www.emdatabank.org
Nucleic Acid Database (NDB) [30]	Nucleic acid structure search and analysis tools	http://ndbserver.rutgers.edu
BioIsis [31]	SAS data deposition	http://www.bioisis.net/
Protein Ensemble Database (pe-DB) [32]	NMR and SAXS data deposition for intrinsically disordered proteins	http://pedb.vib.be
Structural annotation resources		
SCOP and SCOP2 [34, 36]	Structural classification	http://scop.mrc-lmb.cam.ac.uk/scop/
SCOPe [35]	Structural classification	http://scop.berkeley.edu
CATH [37]	Structural classification	http://www.cathdb.info
SUPERFAMILY [39]	Structural alignment and annotation	http://supfam.cs.bris.ac.uk/SUPERFAMILY/
Gene3D [40]	Structural alignment and extended annotation	http://gene3d.biochem.ucl.ac.uk/Gene3D/
Genome3D [41]	Structural alignment and extended annotation	http://genome3d.eu/
ArchDB [42]	Structural classification of loops	http://sbi.imim.es/archdb/
MobiDB [43]	Intrinsically disordered proteins database	http://mobidb.bio.unipd.it
DisProt [44]	Intrinsically disordered proteins database	http://www.disprot.org

URL web addresses are provided, and brief summaries describe the scope of each database

The RCSB PDB Structure Summary page of an entry lists the previously mentioned annotations. Further primary and derived data such as secondary structure depiction, detailed methods information, sequence and structure similarity, protein domain and biological annotations, and many other data are arranged over ten additional tabs to give a complete overview of the macromolecules observed in each structure entry. Additional residue-level and domain-level annotations are retrieved from UniProt and presented in a new “protein feature” view to show the structure–function relationships. Several molecular viewers (Jmol/JSmol, Simple Viewer, Protein Workshop) are integrated to visualize and maneuver the structure within the web browser. RCSB PDB’s Ligand Explorer also allows for detailed visualization of derived protein–ligand annotations, such as involved side chains, and displays calculated bond lengths and interatomic distances [19]. Several web-based structure analysis tools reduce the need for additional software. The Protein Comparison Tool gives access to six algorithms to calculate pairwise sequence or structure alignments [20]. A Drug and Drug Mapping tool combines PDB structural data with *DrugBank* [21] annotations to locate all structures that are bound to pharmaceuticals or are known to act as drug targets.

RCSB PDB also hosts a large education and outreach corner called “PDB-101” that contains hands-on structure-related activities, educational posters, and structural biology tutorials. It also hosts over a decade’s worth of *Molecule of the Month* essays that teach a general audience how the shapes of biologically important macromolecules and their annotated attributes (such as a residue’s role in a catalytic triad or pore selectivity filter) explain their function in living systems.

The *PDBe* provides access to the same PDB archive but has developed different tools to find structures and group them by annotation. Upon entering a PDB ID of interest, it gives “one-click” access to the structural entry and their annotations in *PDBe Atlas*, the ability to download the files, predict protein interfaces and quaternary assembly of the protein using *PDBePISA* [22], find similarly folded structures using *PDBeFold/SSM* [23], or find certain structural motifs or binding sites within each structural chain via the tool *PDBeMotif* [24]. They have also developed a new structure browser based on GO annotations [6]. The *PDBe* is an active contributor to the SIFTS annotation and mapping project, which works to integrate taxonomy, protein sequences, structure, and function [25].

The *PDBj* provides similar structure search, visualization, and analysis tools, and it is also the only member site that supports browsing in additional languages such as Japanese, simplified/traditional Chinese, and Korean. The *SeSAW* tool identifies functionally or evolutionarily conserved motifs in protein structures by locating sequence and structural similarities and annotating these

at the residue level [26]. It has also built tools for bioinformaticians and, in collaboration with the RCSB PDB group, has developed an XML version of the PDB archive (PDBML) [7] and an extended version (PDBMLplus) with additional curated annotations that can be searched using the PDBj Mine application [27].

The *BMRB* became a member of the wwPDB in 2006 [28]. It is a repository that collects NMR data from any experiment, not only from those focused on determining structures. It captures the assigned chemical shifts, coupling constants, and peak lists for a variety of biological macromolecules (even small ones excluded from the PDB). It also contains derived annotations such as hydrogen exchange rates, p*K*_a values, and relaxation parameters that explain biochemical processes in real time. BMRB also collects the NMR restraints for PDB entries, time domain spectral data, and NMR data on hundreds of metabolites and standard compounds [8].

Larger biological assemblies are increasingly studied via three-dimensional electron microscopy (3DEM) methods; these data are available from the *EMDataBank* resource, a joint project of the National Center for Macromolecular Imaging (NCMI), PDBe, and the RCSB [29]. Both maps and model coordinates fitted into the maps are accessible via EMDatabank services. EMDatabank annotations include sample description, sample components, stoichiometry, specimen preparation, details about the imaging experiment, image processing, reconstruction method, resolution, fitting description, and source PDB IDs used in fitting the map. Since 3DEM structures are typically large and complex biological assemblies such as viruses and ribosomes, EMDatabank's sample description annotation permits hierarchical description of the complex components, with the capability to drill down to entity level. Search capabilities are provided at the EMDatabank site.

The *Nucleic Acid Database* (NDB) provides access to information about three-dimensional nucleic acid structures and their complexes [30]. In addition to principal coordinate data, the NDB contains nucleic acid specific annotation and derived structural and geometric data. Nucleic acid structures are deposited to the wwPDB since they are biological macromolecules, and as such, many of the same annotations such as coordinate data, experimental parameters, identifiers, structural description, and data collection and refinement details are captured. Nucleic acid specific annotations such as conformation type, secondary structure information, nucleic acid type, and also protein/drug binding partner information are curated and stored in an Encapsulated Resource File (ERF) that is searchable from the NDB website. Due to the recent growth of RNA biology, the NDB was recently redesigned to include additional derived information on RNA structural features such as pairwise nucleotide interactions, base-stacking interactions, base--phosphate interactions, sequence and

geometry-based equivalence classes containing similar structures, new clustering of non-redundant RNA structure sets, and an “atlas” of representative RNA 3-D motifs extracted from the structures. Users can then search the NDB structures using the aforementioned annotations as constraints, create reports, and download them for further analysis.

Another source of protein structure information is the emerging field of Small Angle Scattering (SAS). These methods are growing in popularity because most proteins are not solely comprised of globular domains; many functional sites are known to be on transiently structured or intrinsically disordered portions of the protein that cannot be visualized by X-ray crystallography or are too large for NMR studies. There are two repositories that currently collect such data: *BioIsis* (SAXS) [31] and the *Protein Ensembles Database* (pe-DB) (NMR and SAS) [32]. Currently, their annotations are limited to capturing experimental parameters and mapping identifiers to UniProt for functional information, but active communities are in the process of further developing their annotation dictionaries [33].

4 Other Structural Annotation Resources

In the early days of protein structure determination, researchers noted that polypeptides would assemble secondary structure elements into regular 3-D patterns (folds) that could be categorized. After many more protein structures were determined, new structures were composed of different combinations of these folds rather than new folds. Those interested more in the biological aspects created resources focused on specific protein superfamilies. This section first describes the databases that classify and annotate the tertiary arrangement of the proteins. It will also briefly summarize the resources that leverage this tertiary information to align similar structures and cluster them. The URL addresses are listed in Table 1.

The Structural Classification of Proteins (*SCOP*) database classifies structures in the PDB in a hierarchical fashion: “family,” “superfamily,” “common fold,” and “class” [34]. A *family* is determined by sequence similarity. A *superfamily* is a set of families with similar structure and function. Families and superfamilies with the same arrangement of secondary structures and connectivity have the same *common fold*. The types of secondary structures in these folds are *classes* (all alpha helix, all beta sheet, alpha-beta, etc.). This type of analysis connects structures with possible evolutionary relationships. One limitation of SCOP is that the current version (1.75) was released in 2009 and contained approximately one-third of today’s PDB archive. A second SCOP group addressed this gap by releasing SCOP-extended (*SCOPe*) that classifies structures added

to the archive since 2009 in an automated fashion [35]. It currently contains 195,523 domains, 4,720 families, 1,978 superfamilies, and 1,205 folds from 68,957 PDB entries (as of the Sept 2014 release). Further review of the data showed that the original hierarchy was too simple, and a new database SCOP2 was released in 2014 [36]. The proteins are still annotated according to their structural and thus inferred evolutionary relationships, but due to the increase in structure data, the tree-like hierarchy had to be expanded to a complex network of nodes. The relationships are now organized into four new categories: protein type, evolutionary events, structural classes, and protein relationships, with mostly new content and definitions and some carryover content from SCOP. The SCOP2 prototype currently encompasses only a select portion of the PDB archive, but this is expected to expand once it is released.

The CATH database uses a similar classification scheme based on class (C), architecture (A), topology (T), and homologous superfamilies (H) [37]. *Class* describes the secondary structure content as in SCOP. *Architecture* defines the description of the arrangement of these secondary structures without consideration of the connectivities. *Topology* is equivalent to fold in SCOP. Finally, *homologous superfamilies* contain all folds with a similar function. In its current release (v4.0), it has annotated 235,858 CATH domains, and 2,738 CATH Superfamilies, from an analysis of ~70,000 PDB entries that were in the archive in 2013. Functional subclassifications (FunFams) have been added to the homologous superfamily layer so that functional divergence within a superfamily could be better understood [38].

Structural annotations such as those curated by CATH and SCOP are being leveraged to help close the sequence-structure gap. *SUPERFAMILY*, part of the SCOP family of databases, organizes structures into evolutionarily related groups to promote the annotation of under-characterized protein [39]. *Gene3D*, part of the CATH family of databases, plays a similar role in assigning CATH domains to gene products of unknown structure [40]. The *Genome3D* project is a collaborative project between SCOP, CATH, and the leading structure prediction resources to extend structural domain and 3-D structure information to more model genomes. In the process, it will serve as the first official mapping between SCOP and CATH [41].

The globular fold is not the only functional interface for proteins, as many binding sites appear on variable loops between α -helices and β -strands. The *ArchDB* classification database classifies loops extracted from structures in the PDB archive [42]. These loops are annotated based on their length, their ϕ and ψ backbone dihedral angles, the type of flanking secondary structures, the distance between the loop ends, and how the loop is supported (braced) by nearby secondary structure elements. The current

version of ArchDB Mar 2014 contains annotations for 306,726 loops spanning 10 different loop types. These data are helpful for the theoretical modeling and protein design communities.

Sometimes the structural classification is that there is—no—shape, with portions of the protein adopting a random coil configuration. The new *MobiDB* database contains the most comprehensive data on intrinsically disordered proteins [43]. *MobiDB* contains 600+ manually curated entries from *DisProt* [44], extends the data with additional evidence of unresolved residues from the PDB (both X-ray and NMR data), adds disordered region predictions from nine bioinformatics resources, and overlays biologically relevant annotations within the regions regarding tissue localization, function, and residue-level annotations from UniProt.

The reason macromolecular structures are annotated with supporting or derived information is to understand the molecular basis of protein function. As a result, there are as many specialty structure annotation databases as there are protein families or structure–function attributes. It is beyond the scope of this paper to cover all of them, so we direct the reader to the Nucleic Acids Research online catalog of protein structure resources at <http://www.oxfordjournals.org/nar/database/subcat/4/14>. This second portion, therefore, will focus only on two major areas of current research: protein–protein interaction databases and membrane–protein annotation databases. The URL addresses for the resources described herein are listed in Table 2.

A protein is rarely isolated in the cell and, sometimes, self-associates or binds to other subunits to become functional. This is not always evident in a structure entry. While not strictly annotation databases, there are many tools that derive these essential 3-D data (and thus deriving new annotations). *PDBePISA* can determine the quaternary assembly of macromolecules derived from calculations made upon the molecule's surface or observed interfaces [45]. The *Dictionary of Interfaces in Proteins* (DIP) is a data bank of complementary molecular surface patches and is meant to enable molecular recognition research [46]. Other bioinformatics tools like *ProtBuD* [47] analyze the surfaces of proteins in the PDB archive to review interactions seen in the crystallographic lattice and to equate them to the likelihood of being the correct biological assembly. A new database of three-dimensional interacting domains (*3did*) annotates binding modes observed in domain–domain interactions and also in domain–peptide interactions from the PDB archive [48]. It has identified 8,944 interactions and classified them into 521 motifs and is updated twice a year.

Another area of intense research is membrane proteins, since they play key roles in controlling the processes of life. Both the Membrane Proteins of Known Structures database (*mpstruc*) [49] and the Protein Data Bank of Transmembrane Proteins (*PDBTM*)

Table 2
Continued list of other annotation databases, data aggregators, and community-driven annotation resources selected for this review

Other structural annotations databases PDBePISA [45]	Intermolecular interactions and quaternary structure	http://pdbe.org/pisa
Dictionary of Interfaces in Proteins (DIP) [46]	Protein-protein interactions	http://dip.doe-mbi.ucla.edu/dip/Main.cgi
ProtBuD [47]	Known and predicted protein interactions	http://dunbrack.fccc.edu/ProtBuD/
3did [48]	Domain-domain and domain-peptide interactions	http://3did.irbbarcelona.org/
Membrane Proteins of Known Structure (mpstruc) [49]	Membrane protein structure search and annotation	http://blanco.biomol.uci.edu/mpstruc/listAll/list
Protein Data Bank of Transmembrane Proteins (PDBTM) [50]	Membrane protein structure search and annotation	http://pdbtm.enzim.hu/
GPCRDB [52]	G protein-coupled receptor protein structure search and annotation	http://www.gpcr.org/7tm/
Transporter Classification Database [53]	Classification of membrane transport proteins	http://www.tcdb.org
Data aggregators		
Structural Biology Knowledgebase (SBKB) [54]	Structure-centric data aggregator and search tool for sites below	http://sbkb.org
PSI Protein Model Portal [55]	Theoretical models data aggregator	http://www.proteinmodelportal.org
TargetTrack [56]	Protein Structure Initiative	http://sbkb.org/tt/
	Structural target registration and experimental history database	
KB-Rank Search tool [58]	Biomedical annotation structure search	http://protein.tcmedc.org
PSI Technology Portal [59]	Catalog of PSI laboratory technologies and methods	http://technology.skbk.org/portal/
PSI Publications Portal [54]	Catalog of articles published by PSI	http://olenka.med.virginia.edu/psi/
Community-driven annotation		
Proteopedia [61]	Community annotation of structures	http://proteopedia.org
TOPSAN [64]	Community annotation of structural genomics structures	http://www.topsan.org

URL web addresses are provided, and brief summaries describe the scope of each database

[50] annotate entries with tertiary structure information along with their orientation, topology, and assembly in the membrane. Historically, mpstruc was manually curated by biological superfamily through reports from the literature, but a recent collaboration with *OMPdb* [51] and the RCSB extended the list of membrane proteins to nearly 1,700 entries. The GPCR database (*GPCRdb*) focuses only on the G protein-coupled receptor superfamily, an important drug target. It integrates residue-level mutation and ligand binding data with their sequences and structures when available [52]. The Transporter Classification Database (*TCDB*) uses a five-level classification scheme to organize phylogenetic and functional information together [53]. The classification includes transporter class, subclass (such as energy source to drive transport), transporter family/superfamily, subfamily, and the range of substrates transported, and also includes filters for human transporters and transporters connected to disease states. If an experimental structure is not available, the TCDB link to external tools that can look for orthologous protein structures or predict transmembrane segments.

5 Data Aggregators

The volume of structural and biological data continues to expand rapidly, and it typically spans over several orders of resolution from molecular mechanisms to cellular phenotypes and tissues. This makes it difficult for non-bioinformaticians to find remote connections between structure and function. Large bioinformatics programs like NCBI (Entrez) and EBI (EBI Search) have search portals that will search all of their underlying databases and return an inventory list of related but disjointed entries. This section describes a structure-centric data aggregator the *Structural Biology Knowledgebase* (*SBKB*), which accesses over one hundred annotation databases. The URL addresses for the resources described in this section are found in Table 2.

The SBKB was established as a scientific portal to facilitate research design and analysis for a wide variety of biological systems [54]. It serves as a single resource that integrates structure, sequence, and functional annotations, as well as technical information regarding protein production and structure determination. Researchers can search the SBKB by sequence, PDB ID, or UniProt accession code, and they receive an up-to-date list of matching 3-D experimental structures from the PDB; prebuilt theoretical models federated by the *Protein Model Portal* [55]; annotations from 100+ open genomic, protein, structural, and functional resources; Protein Structure Initiative (PSI) structural genomics target histories and protocols from *TargetTrack* [56]; and ready-to-use DNA clones from *DNASU* [57]. It is also

possible to find structures by text according to functional and disease relevance (*KB-Rank tool*) [58] or find related technologies and publications from the *PSI Technology Portal* [59] and *PSI Publications Portal*, respectively. Interactive tools such as real-time theoretical modeling and biophysical parameter prediction also enhance the understanding of proteins that are not yet well characterized. A Functional Sleuth viewer collects a list of under characterized proteins that require further study, updated on a weekly basis.

To act as a research concierge, the SBKB has created five “Hubs” that collect practical information about Structural Targets, Modeling, Membrane Proteins, Methods, and “Structure, Sequence, and Function” from its portal modules and the World Wide Web. Select examples include the Membrane Protein Structure Hub which has links to methods for membrane protein production, the latest results from the Protein Structure Initiative’s Human Transmembrane Proteome Coverage project [60], and several public membrane protein structure resources. The “Structure, Sequence and Function” Hub lists all biological resources used during an SBKB query and provides a summary of the search and analysis features for each resource. We refer the reader to this Hub for information on additional resources that could not be included in this review.

6 Community-Driven Structure Annotation Projects

Due to the success of open annotation platforms such as Wikipedia, some groups are taking the same collaborative approach to gather expert information and discussion on scientific topics, including protein structures. This final section will describe recent community annotation efforts for protein structure annotation. The URL addresses for the resources mentioned in this section are found in Table 2.

Proteopedia is web-based structure “wiki” that visualizes molecules and their highlighted structural or functional properties [61]. It allows any registered user to edit any page, and added hyperlinks in the narrative text can trigger the molecular viewer to rotate displaying the highlighted structural feature. Every PDB entry has a page, which at first contains automatically retrieved text from multiple sources such as *ConSurf*, which identifies functional regions in proteins [62], and *OCA*, a structure–function browser [63]. Higher tier articles called “topic pages” describe the protein or protein families and link to individual entry pages. Users can also create the view and export it into Microsoft PowerPoint for teaching purposes. Author-contributed and peer-reviewed Proteopedia pages are also indexed in PubMed.

The Open Protein Structure Annotation Network (*TOPSAN*) is a collaborative annotation platform focused on collecting annotations on structures solved by high-throughput structural genomics efforts [64]. In many cases, their proteins were selected and structurally determined before any functional information was available, so annotating the structure with value-added information presents a challenge. Similar to Proteopedia, a registered contributor can edit a page and add author-created views of the structure and supplemental data in the form of file attachments. The same group at the Joint Center for Structural Genomics has hosted Domain of Unknown Function (DUF) “jamborees” where top bioinformatics experts collaborate to annotate protein families that are listed as having unknown functions; the effort resulted in the discovery of new Pfam domains, some of which have since been included in Pfam.

7 Getting the Most Value Out of Value-Added Annotations

All structural resources have been set up with a common goal—to understand the complex relationship between protein sequence, structure, and function. As indicated in this review, some annotations are experimentally determined, while other value-added annotations are derived or inferred by homology. Rather than keeping the data in silos, all biological (including structural) databases are beginning to integrate cross-referenced annotations from the other resources to make the sequence, structure, and function connections more evident. However, the origins of the annotations (primary vs. derived) are sometimes buried in documentation (or worse, ignored) and then annotations that were originally inferred by homology are suddenly understood to be fact. There can also be a lag in time between the release of a new study and its incorporation into the databases. This leads to a cyclical propagation of errors, which can linger when some databases do not release updates as often as UniProt (approx. monthly) or PDB (weekly). There are plenty of articles that point out errors in the reference databases [65–67], including a recent case study of the repair of an incorrect annotation by the UniProtKB Consortium [68]. To the credit of the annotation databases, there is an immense amount of sequence data that needs to be covered and they are keen to quickly repair inconsistencies. But as data-users, we must be aware of the caveats that are implied in derived data, and to always check (and cite) the original source, so that validated science is always propagated.

Acknowledgements

The authors are grateful to B. Coimbatore Narayanan, G. Gabanyi, C. Lawson, E. Peisach, M. Quesada, M. Sekharan, J. Westbrook, and C. Zardecki for helpful discussions and reviews during the preparation of this work. The RCSB PDB (H.M.B.) is supported by the National Science Foundation (NSF DBI 1338415), the Department of Energy, the National Library of Medicine, the National Cancer Institute, the National Institute of Neurological Disorders and Stroke, and the National Institute of Diabetes and Digestive and Kidney Diseases. The SBKB (H.M.B./M.G.) is supported by a grant from the National Institute of General Medical Sciences of the National Institutes of Health (U01 GM093324).

References

1. Benson DA, Clark K, Karsch-Mizrachi I et al (2014) GenBank. *Nucleic Acids Res* 42: D32–D37. doi:[10.1093/nar/gkt1030](https://doi.org/10.1093/nar/gkt1030)
2. The UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 42:D191–D198. doi:[10.1093/nar/gkt1140](https://doi.org/10.1093/nar/gkt1140)
3. Berman HM, Westbrook JD, Feng Z et al (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242
4. Berman HM, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 10:980. doi:[10.1038/nsb1203-980](https://doi.org/10.1038/nsb1203-980)
5. The Protein Data Bank (1971) Protein Data Bank. *Nat New Biol* 233:223. doi:[10.1038/newbio233223b0](https://doi.org/10.1038/newbio233223b0)
6. Gutmanas A, Alhroub Y, Battle GM et al (2014) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res* 42:D285–D291. doi:[10.1093/nar/gkt1180](https://doi.org/10.1093/nar/gkt1180)
7. Kinjo AR, Suzuki H, Yamashita R et al (2012) Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. *Nucleic Acids Res* 40:D453–D460. doi:[10.1093/nar/gkr811](https://doi.org/10.1093/nar/gkr811)
8. Ulrich EL, Akutsu H, Doreleijers JF et al (2008) BioMagResBank. *Nucleic Acids Res* 36:D402–D408. doi:[10.1093/nar/gkm957](https://doi.org/10.1093/nar/gkm957)
9. Read RJ, Adams PD, Arendall WB III et al (2011) A new generation of crystallographic validation tools for the Protein Data Bank. *Structure* 19:1395–1412. doi:[10.1016/j.str.2011.08.006](https://doi.org/10.1016/j.str.2011.08.006)
10. Montelione GT, Nilges M, Bax A et al (2013) Recommendations of the wwPDB NMR Validation Task Force. *Structure* 21:1563–1570. doi:[10.1016/j.str.2013.07.021](https://doi.org/10.1016/j.str.2013.07.021)
11. Henderson R, Sali A, Baker ML et al (2012) Outcome of the first electron microscopy validation task force meeting. *Structure* 20:205–214. doi:[10.1016/j.str.2011.12.014](https://doi.org/10.1016/j.str.2011.12.014)
12. The wwPDB Consortium (2013) New wwPDB X-ray structure validation reports support depositors, journal editors and referees. http://wwpdb.org/news/news_2013.html#02-August-2013. Accessed 1 Mar 2014
13. The wwPDB Consortium (2013) Deposition and release of PDB entries containing large structures. http://wwpdb.org/news/news_2013.html#22-May-2013. Accessed 1 Mar 2014
14. The wwPDB Annotation Staff (2014) wwPDB processing procedures and policies document: section B: wwPDB policies. <http://www.wwpdb.org/policy.html>. Accessed 1 Mar 2014
15. Westbrook JD, Fitzgerald PMD (2009) Chapter 10 The PDB format, mmCIF formats, and other data formats. In: Bourne PE, Gu J (eds) *Structural bioinformatics*, 2nd edn. Wiley, Hoboken, NJ, pp 271–291
16. Dutta S, Dimitropoulos D, Feng Z et al (2014) Improving the representation of peptide-like inhibitor and antibiotic molecules in the Protein Data Bank. *Biopolymers* 101:659–668. doi:[10.1002/bip.22434](https://doi.org/10.1002/bip.22434)
17. Quesada M, Westbrook J, Oldfield T et al (2011) The wwPDB common tool for deposition and annotation. *Acta Cryst*, C403–C404
18. Rose PW, Bi C, Bluhm WF et al (2013) The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res* 41:D475–D482. doi:[10.1093/nar/gks1200](https://doi.org/10.1093/nar/gks1200)

19. Moreland JL, Gramada A, Buzko OV et al (2005) The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications. *BMC Bioinformatics* 6:21. doi:[10.1186/1471-2105-6-21](https://doi.org/10.1186/1471-2105-6-21)
20. Prlic A, Bliven S, Rose PW et al (2010) Pre-calculated protein structure alignments at the RCSB PDB website. *Bioinformatics* 26:2983–2985. doi:[10.1093/bioinformatics/btq572](https://doi.org/10.1093/bioinformatics/btq572)
21. Knox C, Law V, Jewison T et al (2011) DrugBank 3.0: a comprehensive resource for ‘omics’ research on drugs. *Nucleic Acids Res* 39:D1035–D1041. doi:[10.1093/nar/gkq1126](https://doi.org/10.1093/nar/gkq1126)
22. Krissinel E, Henrick K (2005) Detection of protein assemblies in crystals. In: Berthold MR, Glen R, Diederichs K, Kohlbacher O, Fischer I (eds) *Computational life sciences. First international symposium, CompLife 2005, Konstanz, Germany, September 25–27, 2005, Proceedings*. Springer-Verlag, Berlin, pp 163–174
23. Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr* 60:2256–2268. doi:[10.1107/S0907444904026460](https://doi.org/10.1107/S0907444904026460)
24. Golovin A, Henrick K (2009) Chemical substructure search in SQL. *J Chem Inf Model* 49:22–27. doi:[10.1021/ci8003013](https://doi.org/10.1021/ci8003013)
25. Velankar S, Dana JM, Jacobsen J et al (2013) SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res* 41:D483–D489. doi:[10.1093/nar/gks1258](https://doi.org/10.1093/nar/gks1258)
26. Standley DM, Yamashita R, Kinjo AR et al (2010) SeSAW: balancing sequence and structural information in protein functional mapping. *Bioinformatics* 26:1258–1259. doi:[10.1093/bioinformatics/btq116](https://doi.org/10.1093/bioinformatics/btq116)
27. Kinjo AR, Yamashita R, Nakamura H (2010) PDBj Mine: design and implementation of relational database interface for Protein Data Bank Japan. *Database (Oxford)* 2010:baq021. doi:[10.1093/database/baq021](https://doi.org/10.1093/database/baq021)
28. Markley JL, Ulrich EL, Berman HM et al (2008) BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *J Biomol NMR* 40:153–155
29. Lawson CL, Baker ML, Best C et al (2011) EMDatabank.org: unified data resource for CryoEM. *Nucleic Acids Res* 39:D456–D464. doi:[10.1093/nar/gkq880](https://doi.org/10.1093/nar/gkq880)
30. Coimbatore Narayanan B, Westbrook J, Ghosh S et al (2014) The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Res* 42:D114–D122. doi:[10.1093/nar/gkt980](https://doi.org/10.1093/nar/gkt980)
31. Hura GL, Menon AL, Hammel M et al (2009) Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). *Nat Methods* 6:606–612. doi:[10.1038/nmeth.1353](https://doi.org/10.1038/nmeth.1353)
32. Varadi M, Kosol S, Lebrun P et al (2014) pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res* 42:D326–D335. doi:[10.1093/nar/gkt960](https://doi.org/10.1093/nar/gkt960)
33. Trewella J, Hendrickson WA, Sato M et al (2013) Meeting report of the wwPDB small-angle scattering task force: data requirements for biomolecular modeling and the PDB. *Structure* 21:875–881
34. Andreeva A, Howorth D, Brenner SE et al (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 32:D226–D229
35. Fox NK, Brenner SE, Chandonia JM (2014) SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res* 42:D304–D309. doi:[10.1093/nar/gkt1240](https://doi.org/10.1093/nar/gkt1240)
36. Andreeva A, Howorth D, Chothia C et al (2014) SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res* 42:D310–D314. doi:[10.1093/nar/gkt1242](https://doi.org/10.1093/nar/gkt1242)
37. Cuff AL, Sillitoe I, Lewis T et al (2009) The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res* 37:D310–D314. doi:[10.1093/nar/gkn877](https://doi.org/10.1093/nar/gkn877)
38. Sillitoe I, Cuff AL, Dessailly BH et al (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res* 41:D490–D498. doi:[10.1093/nar/gks1211](https://doi.org/10.1093/nar/gks1211)
39. Wilson D, Madera M, Vogel C et al (2007) The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res* 35:D308–D313
40. Lees JG, Lee D, Studer RA et al (2014) Gene3D: multi-domain annotations for protein sequence and comparative genome analysis. *Nucleic Acids Res* 42:D240–D245. doi:[10.1093/nar/gkt1205](https://doi.org/10.1093/nar/gkt1205)
41. Lewis TE, Sillitoe I, Andreeva A et al (2013) Genome3D: a UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains. *Nucleic Acids Res* 41:D499–D507. doi:[10.1093/nar/gks1266](https://doi.org/10.1093/nar/gks1266)

42. Bonet J, Planas-Iglesias J, Garcia-Garcia J et al (2014) ArchDB 2014: structural classification of loops in proteins. *Nucleic Acids Res* 42:D315–D319. doi:[10.1093/nar/gkt1189](https://doi.org/10.1093/nar/gkt1189)
43. Di Domenico T, Walsh I, Martin AJ et al (2012) MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics* 28:2080–2081. doi:[10.1093/bioinformatics/bts327](https://doi.org/10.1093/bioinformatics/bts327)
44. Sickmeier M, Hamilton JA, LeGall T et al (2007) DisProt: the Database of Disordered Proteins. *Nucleic Acids Res* 35:D786–D793. doi:[10.1093/nar/gkh893](https://doi.org/10.1093/nar/gkh893)
45. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 372:774–797. doi:[10.1016/j.jmb.2007.05.022](https://doi.org/10.1016/j.jmb.2007.05.022)
46. Salwinski L, Miller CS, Smith AJ et al (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res* 32:D449–D451. doi:[10.1093/nar/gkh086](https://doi.org/10.1093/nar/gkh086)
47. Xu Q, Canutescu A, Obradovic Z et al (2006) ProtBuD: a database of biological unit structures of protein families and superfamilies. *Bioinformatics* 22:2876–2882. doi:[10.1093/bioinformatics/btl490](https://doi.org/10.1093/bioinformatics/btl490)
48. Mosca R, Ceol A, Stein A et al (2014) 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res* 42:D374–D379. doi:[10.1093/nar/gkt887](https://doi.org/10.1093/nar/gkt887)
49. Snider C, Jayasinghe S, Hristova K et al (2009) MPEx: a tool for exploring membrane proteins. *Protein Sci* 18:2624–2628. doi:[10.1002/pro.256](https://doi.org/10.1002/pro.256)
50. Kozma D, Simon I, Tusnady GE (2013) PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res* 41:D524–D529. doi:[10.1093/nar/gks1169](https://doi.org/10.1093/nar/gks1169)
51. Tsirigos KD, Bagos PG, Hamodrakas SJ (2011) OMPdb: a database of {beta}-barrel outer membrane proteins from Gram-negative bacteria. *Nucleic Acids Res* 39:D324–D331. doi:[10.1093/nar/gkq863](https://doi.org/10.1093/nar/gkq863)
52. Isberg V, Vroiling B, van der Kant R et al (2014) GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res* 42:D422–D425. doi:[10.1093/nar/gkt1255](https://doi.org/10.1093/nar/gkt1255)
53. Saier MH Jr, Reddy VS, Tamang DG et al (2014) The transporter classification database. *Nucleic Acids Res* 42:D251–D258. doi:[10.1093/nar/gkt1097](https://doi.org/10.1093/nar/gkt1097)
54. Gabanyi MJ, Adams PD, Arnold K et al (2011) The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. *J Struct Funct Genomics* 12:45–54. doi:[10.1007/s10969-011-9106-2](https://doi.org/10.1007/s10969-011-9106-2)
55. Haas J, Roth S, Arnold K et al (2013) The Protein Model Portal—a comprehensive resource for protein structure and model information. *Database* (Oxford) 2013:bat031. doi:[10.1093/database/bat031](https://doi.org/10.1093/database/bat031)
56. Chen L, Oughtred R, Berman HM et al (2004) TargetDB: a target registration database for structural genomics projects. *Bioinformatics* 20:2860–2862. doi:[10.1093/bioinformatics/bth300](https://doi.org/10.1093/bioinformatics/bth300)
57. Seiler CY, Park JG, Sharma A et al (2014) DNASU plasmid and PSI: Biology-Materials repositories: resources to accelerate biological research. *Nucleic Acids Res* 42:D1253–D1260. doi:[10.1093/nar/gkt1060](https://doi.org/10.1093/nar/gkt1060)
58. Julfayev ES, McLaughlin RJ, Tao YP et al (2012) KB-Rank: efficient protein structure and functional annotation identification via text query. *J Struct Funct Genomics* 13:101–110. doi:[10.1007/s10969-012-9125-7](https://doi.org/10.1007/s10969-012-9125-7)
59. Gifford LK, Carter LG, Gabanyi MJ et al (2012) The Protein Structure Initiative Structural Biology Knowledgebase Technology Portal: a structural biology web resource. *J Struct Funct Genomics* 13:57–62. doi:[10.1007/s10969-012-9133-7](https://doi.org/10.1007/s10969-012-9133-7)
60. Pieper U, Schlessinger A, Kloppmann E et al (2013) Coordinating the impact of structural genomics on the human alpha-helical transmembrane proteome. *Nat Struct Mol Biol* 20:135–138. doi:[10.1038/nsmb.2508](https://doi.org/10.1038/nsmb.2508)
61. Prilusky J, Hodis E, Canner D et al (2011) Proteopedia: a status report on the collaborative, 3D web-encyclopedia of proteins and other biomolecules. *J Struct Biol* 175:244–252. doi:[10.1016/j.jsb.2011.04.011](https://doi.org/10.1016/j.jsb.2011.04.011)
62. Ashkenazy H, Erez E, Martz E et al (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* 38:W529–W533. doi:[10.1093/nar/gkq399](https://doi.org/10.1093/nar/gkq399)
63. Prilusky J (1996) OCA, a browser-database for protein structure/function. <http://oca.weizmann.ac.il>. Accessed 1 Mar 2014
64. Krishna SS, Weekes D, Bakolitsa C et al (2010) TOPSAN: use of a collaborative environment for annotating, analyzing and disseminating data on JCSG and PSI structures. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 66:1143–1147
65. Zheng H, Chordia MD, Cooper DR et al (2014) Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server. *Nat Protoc* 9:156–170. doi:[10.1038/nprot.2013.172](https://doi.org/10.1038/nprot.2013.172)
66. Richardson CR, Luo QJ, Gontcharova V et al (2010) Analysis of antisense expression by

- whole genome tiling microarrays and siRNAs suggests mis-annotation of Arabidopsis orphan protein-coding genes. PLoS One 5:e10710. doi:[10.1371/journal.pone.0010710](https://doi.org/10.1371/journal.pone.0010710)
67. Schnoes AM, Brown SD, Dodevski I et al (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5:e1000605. doi:[10.1371/journal.pcbi.1000605](https://doi.org/10.1371/journal.pcbi.1000605)
68. Poux S, Magrane M, Arighi CN et al (2014) Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data. Database (Oxford) 2014:bau016. doi:[10.1093/database/bau016](https://doi.org/10.1093/database/bau016)

Chapter 2

PiMS: A Data Management System for Structural Proteomics

Chris Morris

Abstract

PiMS (Protein Information Management System) is a laboratory information management system for protein scientists. It enables researchers to enter data, track samples, and report results during the production of recombinant proteins for structural and functional applications. PiMS is the only custom LIMS for protein production, recording data from the selected target to the sample of soluble protein. The xtal-PiMS extension supports crystallogenesis and has recently been extended to support crystal fishing and crystal treatment. PiMS can be configured to match local working methods by defining protocols. These are used to provide templates for recording details of the experiments. PiMS will continue to be developed in response to the needs of users to provide a unified and extensible set of software tools for protein sciences. The vision for PiMS is that it will become the laboratory standard for protein-related data management. The Science and Technology Facilities Council (STFC) distributes PiMS free to academic users under the Community Model.

Key words LIMS, Laboratory information management system, Research data, Protein expression, Recombinant protein, ELN, Electronic laboratory notebook, Laboratory notebook

1 Introduction

1.1 *The Traditional Approach to Laboratory Information Management*

Scientists record their laboratory activity in order to:

- Be able to publish their results
- Have sufficient records that they and others can repeat their experiments
- Avoid repeating work
- Coordinate work with collaborators
- Report to supervisors [1]

The laboratory notebook, a paper notebook owned by one scientist, has a key role in the traditions and current practice of science, and undergraduate courses include training in how to keep one [2]. In practice, a great deal of information is also recorded

in spreadsheets, MS Word documents, and some existing local databases [1]. Scientists also use Web sites to look up gene sequences, protein sequences, and standard protocols.

Over the last 5 years, there has been a significant convergence in methods used for protein production. Nevertheless, most laboratories implement at least one stage of the protein production pipeline in a special way. Laboratories also vary considerably in their preferred vocabulary, even when they are describing the same things. For example, the words “target” and “construct” have a wide spread of meanings.

The people who read and record laboratory information include:

- Independent researchers, who are responsible for work on one or more targets, from start to end of the pipeline
- Supervised students, who work on one or more targets, from start to end of the pipeline
- Principal investigators, who manage a whole research grant, involving many targets
- Laboratory technicians, who may manage all the work that passes through a specific instrument at a specific stage in the pipeline, e.g., purification columns or may work on one or more targets from start to end of the pipeline
- Archivists or systems administrators, who are responsible for supporting the records

1.2 Limitations of the Traditional Approach

The current system of record keeping is stressed as structural biologists determine more structures each year (Fig. 1), and more complex structures (Fig. 2) [3]. Current record keeping is also inadequate for long-running research projects on difficult targets, which can take longer than a single postdoctoral contract. It is not uncommon to find samples in freezers with no adequate record of what they are or how they were produced, resulting in the waste of the materials and labor used in their production, and the possible loss of a research opportunity [4]. Scientists currently use a variety of independent IT tools, and errors can be introduced exchanging data between these with potentially wasteful consequences.

In the private sector, there is a search for new business models for drug discovery, often involving collaborations across organizational boundaries [5]. Even within academia, collaboration at a distance is now common. It is also increasingly common for some processes to be carried out on a service basis, e.g., crystallogensis or the creation of constructs. Each such service has had to create its own ad hoc mechanisms for reporting back to its clients.

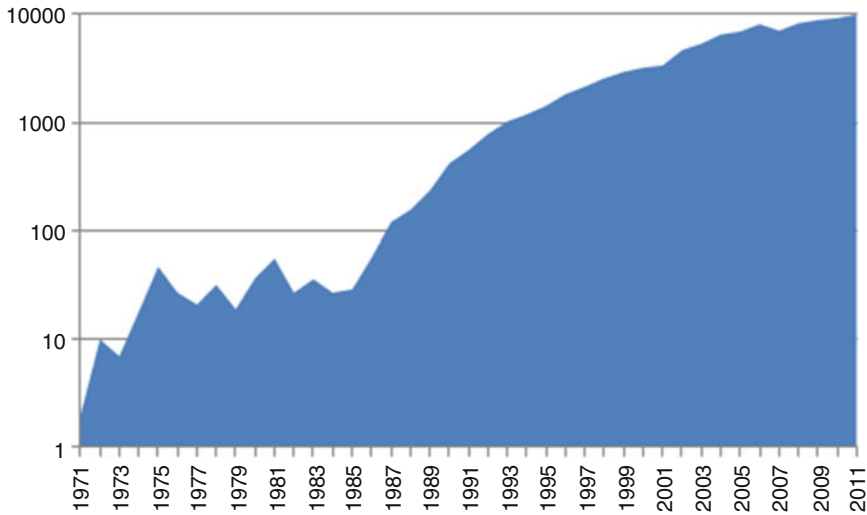


Fig. 1 Yearly accessions to the PDB shown on a log₁₀ scale

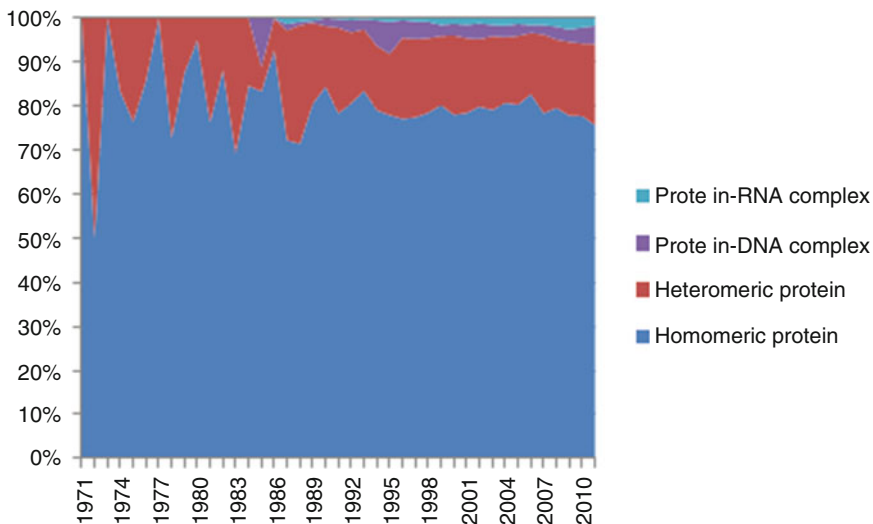


Fig. 2 Accessions to the PDB by year, showing the proportion of single proteins and protein complexes

Standardizing laboratory records will enable increasing amounts of data to be trapped automatically without requiring intervention from scientists and allow records in an electronic form to be archived and shared, facilitating the reproduction of results.

The priorities for the development of PiMS were:

1. A user-friendly system for entering and retrieving records of targets (open reading frames of interest)
2. A user-friendly system for entering and retrieving records of experiments, including experiments in plates

3. Support for primer design and managing constructs
4. Automatically uploading files from instruments, as resources permit

Data management by PiMS and the xtalPiMS extension stops at the pipeline stage when the crystal goes to the synchrotron. Data management after that point is the responsibility of sister projects. Anticipated future developments will involve the integration of bioinformatic tools to support target selection.

1.3 The Design Space for Laboratory Information Management

An Electronic Laboratory Notebook (ELN) provides free form, unstructured data entry. These are suitable for very variable processes. For example, a university chemistry laboratory rarely repeats an experiment. This convenience at the time of data entry comes at the cost that searching the data is not easy.

A Laboratory Information Management System (LIMS) has a database that incorporates the ideas of projects, samples, and standard operating procedures and structures the data entry accordingly. There are many commercial LIMS products, which are designed to support highly reproducible processes. For example, a hospital blood laboratory does the same analysis 100,000s of times a year.

Protein production is in an intermediate case, with some standard steps and also a great deal of variation. The design goal for PiMS was to provide a user experience matching an ELN, along with the benefits of a data structure that understands not only the standard concepts of a LIMS but also the concepts specific to molecular biology, e.g., primers.

The available data management software can be positioned in other dimensions too:

- A collection of MS Word documents provides equal support to every stage of the pipeline, whereas a VectorNTI database supports one specific stage.
- A spreadsheet is essentially private, and sharing it is possible but inconvenient. A wiki is essentially public, and limiting access requires some work. PiMS is in the middle of this range, guaranteeing the privacy of the data entered, with flexibility to allow scientists to join and leave collaborations and share data appropriately.

Table 1 shows the position of some software products and services on two of these dimensions and thereby indicates some of the design choices made in the development of PiMS.

The most important quality attribute for PiMS is reliability. Scientists trust it with their most important asset, experimental data. All the data that scientists think they have entered into PiMS must indeed be saved and be retrievable. We use a variety of testing technologies to ensure this.

Table 1
The design space for data management software, authorisation versus data modeling

	Free form	Structured	Mol. Biol aware
Private	MS Word	MS Excel	VectorNti
Controlled	ConturELN	Nautilus Sugar CRM	PiMS
Public	MediaWiki	–	PDB

The second most important quality attribute is usability. The benefits of PiMS are felt when data is retrieved. The costs are paid earlier, when data is saved. Therefore, data entry must be as painless as possible.

The most important uses of PiMS are as follows:

- Planning experiments: reviewing alternative protocols and previous results, recording protocols, and recording plans for lab experiments
- Recording the results of lab experiments, including experiments carried out in multiwell plates
- Reporting to clients of service operations, and preparing journal publications
- Supervision: keeping track of projects
- Managing health and safety records
- Managing reagent stocks
- Archiving completed research projects

1.4 Sustainability

PiMS is distributed free by STFC for academic use [6]. The PIMS Steering Committee facilitates the coordination of PiMS releases and further development of PIMS to assure future compatibility of PIMS core functionality, and reports annually to the PiMS community.

In the academic sector, funding organizations state the expectation that experimental results are to be made public, and to be archived for an extended period. The first such PiMS dataset has been contributed by the OPPF-UK [7].

Academics have a strong expectation that software will be free at the time of use [4]. They also increasingly expect that it will be of high quality and will be maintained. This is particularly important for data management software—if a structure solution program fails, then you can try another one, but if a data repository fails, years of work can be lost. The average resource required to

maintain software is one full-time equivalent for maintenance for every seven person years development effort [8], and at the time of writing the development of PiMS has taken 45 person years of effort. The funding mechanisms available for research software do not provide continuation funding on this scale, so STFC intends to use the royalties it receives from sales of PiMS, to help support academic users of PiMS. In addition, CCP4 has generously funded the ongoing support of PiMS.

There is a hosted PiMS service (<http://pims.structuralbiology.eu/docs/index.html>), which has been scaled to support all the academic protein production service in Europe. Current resources do not permit a hosted xtalPiMS service.

2 Using PiMS

2.1 *Logging in*

The PiMS administrator will allocate a username and password and will tell you the URL of your PiMS installation. You can then log in to PiMS, using any current browser (e.g., Firefox, Internet Explorer, Safari, or Chrome). You will be allocated to one or more user groups with permissions appropriate for your work—you can ask for a private “lab notebook” within PiMS or share one or more with collaborators.

2.2 *Menus*

There is a menu bar near the top of the screen (Fig. 3). The words in it are the titles of different PiMS menus. If you put your mouse over it, a more detailed menu will drop down. In the screenshot, the user is about to click “New Target.”

2.3 *New Target*

Most PiMS activity begins by identifying a “target,” that is to say an open reading frame of interest. The New Target page allows you to specify a database and an accession code. PiMS will download details for you, including links to other databases.

Figure 4 shows the new page in PiMS with this information. Like most pages in PiMS, it is a part of a “laboratory notebook,” private to a specified set of users. The information in it is organized in a series of boxes, which the user can open or close. Links near the top of the page provide quick access to a number of actions. In this case, the next one is usually “New construct.”

2.4 *New Construct*

A Construct in PiMS is a plan for expression: the design of a primers and a plasmid, to express a full length open reading frame, a truncation, or perhaps parts of more than one open reading frame.

In Fig. 5, the user has chosen full-length expression. After she or he clicks “Design Primers,” the next page will show suggested primer overlap sequences, and then offer a menu of extensions, then the user can click “Save.”

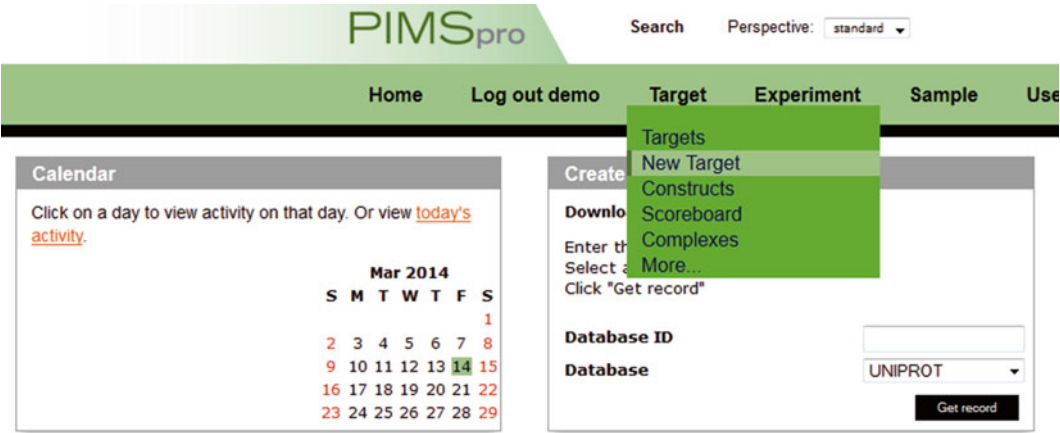


Fig. 3 Screenshot showing the drop-down menus in PiMS



Fig. 4 Screenshot showing the record of target in PiMS with details of DNA and corresponding amino acid sequence with some calculated properties

At the top of the page, a “breadcrumb trail” provides easy navigation back from the construct to the target. A large icon indicates that this record is a Construct (Fig. 6). The key types that PiMS users will become familiar with are Targets, Constructs, Experiments, and Samples.

2.5 Experiments

To record a new experiment, there is a link on the view of a Construct, and there is also a link on the view of a Sample. PiMS highlights in yellow the actions you are most likely to want to perform next: in this case, to record the conditions of the experiment

Targets : OPTIC14032

New Construct: Basic details

Translated Protein Sequence -translated from the Target DNA sequence

Please note that this is translated from the Target DNA sequence.

MCCRAIKHRA QGLVAFAIL LLLAGWVT GAAMPANIAA AMSGPLEVA ICAEGHAATI WLDAEGNEHP APQECRDCPV CHPPALTADP QPQLPAAPGR 100
WLPQATRLAA AQVRGAGREL LVQVRGPPSF SSANTRAVFS PAGDRPSFDA VDPQPMRGI RAIARDARA 169

[Pop-up](#) view of the Target DNA and translated Protein sequence

Basic Details

Construct Id*

OPTIC14032

Target protein start*

1

Target protein end*

169

Lab Notebook

OPPF

Choose one:

Primer design Tm 60

Save and Design Primers >

Primerless construct

Save Construct

Fig. 5 Screenshot of the form for entering details of a new construct in PiMS

Targets : OPTIC12846

Construct: OPPF10000

[New Experiment](#) [Diagram](#) [Delete](#) [Milestones](#) [New SDM Primers](#)

Basic Details

Page Number

Construct Name

Description

OPPF10000

Scientist

Comments

Lab Notebook: OPPF

Forward Primer: OPPF10000F

Full sequence

AAGTTCGTGTTTCAGGGCCCG ATTACTACAAGGGAAGACAT AAATTCAAAGCAGG

Length: 54 Tm °C: 74.7 %GC: 42.6 Molecular Mass: 16,773 [Fasta pop-up](#)

Overlap region

[ATTACTACAAGGGAAGACAT AAATTCAAAGCAGG](#)

Length: 34 Tm °C: 64.7 %GC: 35.3

5'-Extension

[AAGTTCGTGTTTCAGGGCCCG](#)

Reverse Primer: OPPF10000R

Fig. 6 Screenshot showing details of the PCR primers for construct design in PiMS

and any observations, and to record your conclusion about whether it succeeded (Fig. 7). Once you have changed the status to “OK” or “Failed,” the record of the experiment is locked. Only some PiMS users have permission to unlock a page, and the unlocking is

The screenshot displays the PiMS interface for a planned experiment. At the top, the title '10942 Production scale expression' is prominent. Below the title, there are tabs for 'Diagram', 'Copy', 'Delete', 'Not locked', and 'Help'. A calendar widget on the right shows the date '04 April 2013'. The main section is titled 'Basic Details' and contains the following information:

- Name:** 10942 Production scale expression
- Type:** Production scale expression
- Protocol:** OPPF ScaleUp Expression Simple
- Lab Notebook:** Protein 100
- Status:** To be run
- Milestone:** ☐ Production Scale expression achieved
- Start date:** 04 April 2013 12:48:59 GMT+0100
- End date:** 04 April 2013 12:48:59 GMT+0100
- Project:** OPPF10942
- Scientist:** yamini

Below the 'Basic Details' section is the 'Conditions and Results' section, which contains a table with the following data:

Parameter	Value
Selenomethionine Labelled	No
Induction	PB + IPTG
Cells	Rosetta

At the bottom of the interface, there are tabs for 'Reagents' and 'Samples'. A 'Make changes...' button is visible at the bottom right of the 'Basic Details' section.


Fig. 7 Screenshot of planned experiment recorded in PiMS

permanently recorded. By configuring these permissions, PiMS can be part of a working procedure that conforms to CFR21 Part 11, the US Food and Drug Administration guidelines for electronic notebook and electronic signatures. You add notes, links, images, and other files to any page in PiMS (Fig. 8). The view of an experiment or sample provides a link to a “diagram” that summarizes the work done in graphical form. It gives a quick way to navigate back to earlier experiments in the production process (Fig. 9).

2.6 Protocols





A “Protocol” in PiMS is the template for an experiment (Fig. 10). It specifies the types of samples used and produced, the observations that will be made, and the decisions that must be taken in preparation. These “Set up parameters” contain the information that the experimenter will need to reproduce the experiment. Someone from another laboratory would in addition need to know the group’s standard operating procedure. This can be recorded in the “Method” box, or by an attached file.

PiMS is supplied with a range of standard protocols. Nevertheless, the key to making it truly usable is to customize these to the actual working methods of the laboratory. It is natural to ask “what data do we want to record?” but it has proved more helpful to ask “what data will we want to look up?” Groups that use PiMS usually review and simplify their protocols after 1 or 2 years.



Experiments : Protein purification Experiments

11660 protein purification 25/2/13

 [Diagram](#)
 [Copy](#)
 [Delete](#)
[Not locked](#)
 [Help](#)

Mar 2014

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

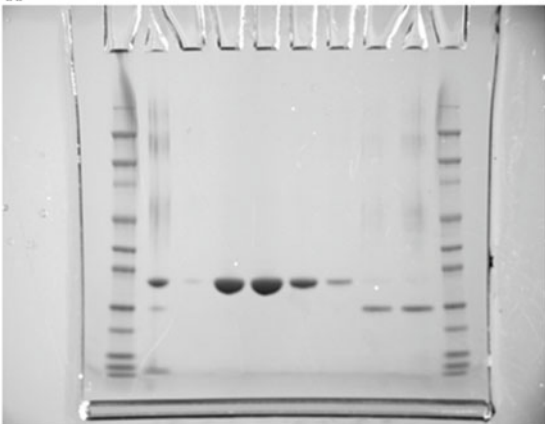
Basic Details

Conditions and Results


Reagents

Samples

Images



14 Mar 2014 15:08:41 GMT

 [Make changes...](#)

Upload a file:

No file selected.

Fig. 8 Screenshot of experimental data (stained protein gel) added as an attachment to an experiment in PiMS

You can link together your protocols into one or more planned workflows, e.g., a standard gene-to-structure workflow, which you assign to most constructs when you design them, and a recovery workflow, which you add if soluble expression fails.

There are many other features in PiMS. The online help files supply more details (Fig. 11).

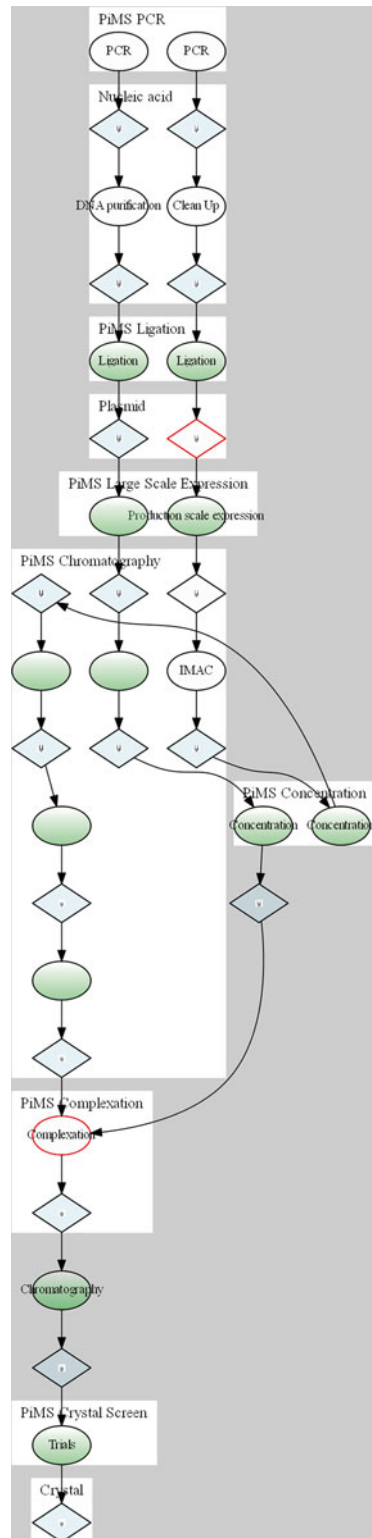


Fig. 9 The diagrammatic summary of a project generated from records in PiMS



PiMS has been designed to support recombinant protein production with the convenience of an ELN at the time of data entry and the benefits of LIMS at the time of data retrieval.

The adoption of PiMS will improve the efficiency of information management by avoiding some existing waste and inefficiency

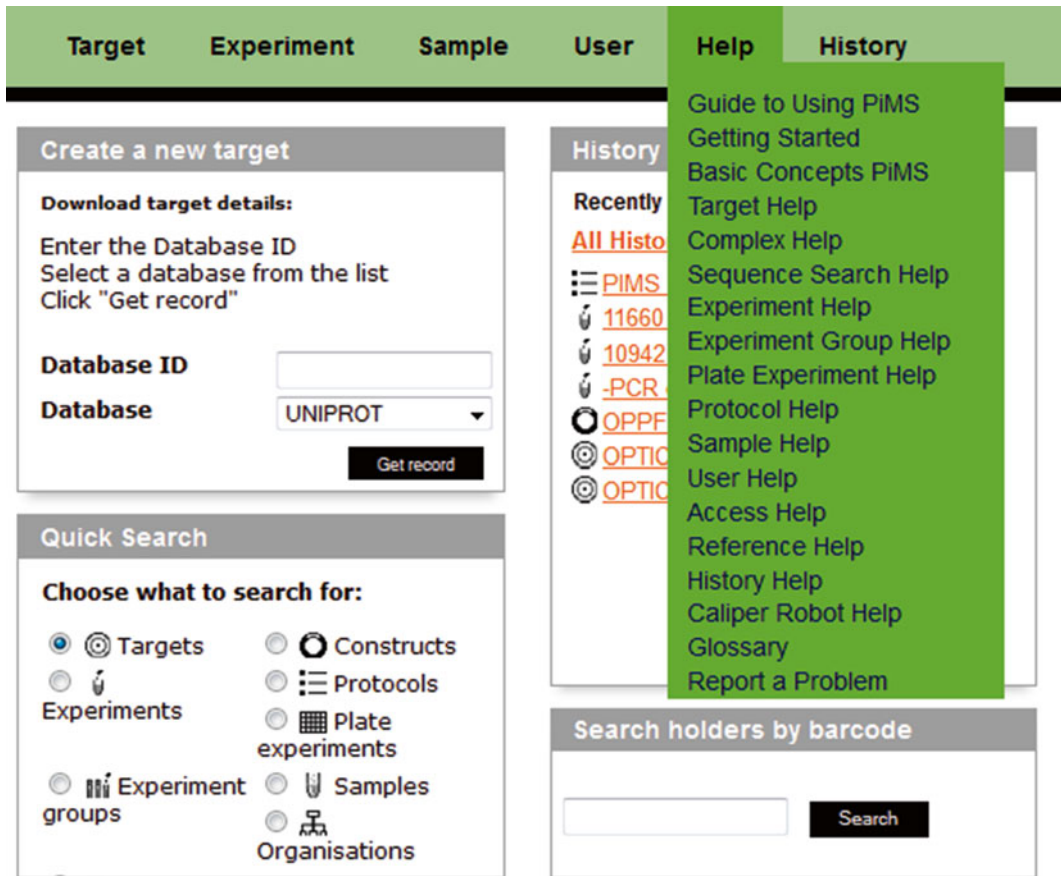


Fig. 11 Screenshot of the help pages in PiMS

in the research process and will help scientists become more productive individually as research teams and as a community. However, the adoption of a LIMS is a significant culture change. To get the full benefits requires a greater level of discipline in the laboratory. It would be expected to facilitate communication between scientists, including the sharing of data in collaborations and oversight by supervisors and funding agencies. Understandably, these changes are not unreservedly welcomed.

The adoption of PiMS requires an investment of time to learn it and adapt the PiMS standard protocols to match local practice (about a person week). Most laboratories that have adopted PiMS later revise downwards their initial decisions about how much data to record. It would usually be better to begin with a minimal level of required data entry, and then increase the standard required as it proves useful to do so.

The productivity of structural biologists has improved over the years because of hard work, better reagents, better instruments, and better software. Now they require an integrated software tool kit, with a consistent user interface and seamless data transfer, from target selection to structure interpretation. PiMS is seen as one component of that future tool kit.

References

1. <http://pims.structuralbiology.eu/docs/Q-2010-01.doc>
2. Caprette D (2006) Guidelines for keeping a laboratory record, Rice University. <http://www.ruf.rice.edu/~bioslabs/tools/notebook/notebook.html>
3. Figures provided by Sameer Velankar, EBI
4. Interviews by author
5. Canady M (2012) From outsourced to open: the continuing evolution of the drug discovery business model. DDW Spring 12
6. <http://pims.structuralbiology.eu/docs/community%20model%20pimsacademicuse-only%20FINAL.docx>
7. <http://pims.structuralbiology.eu:8080/rdf/html/index.html>
8. Jones C (2006) The economics of software maintenance in the twenty first century. <http://www.compaid.com/caiinternet/ezone/capersjones-maintenance.pdf>.

Prediction and Analysis of Intrinsically Disordered Proteins

Marco Punta, István Simon, and Zsuzsanna Dosztányi

Abstract

Intrinsically disordered proteins and protein regions (IDPs/IDRs) do not adopt a well-defined folded structure under physiological conditions. Instead, these proteins exist as heterogeneous and dynamical conformational ensembles. IDPs are widespread in eukaryotic proteomes and are involved in fundamental biological processes, mostly related to regulation and signaling. At the same time, disordered regions often pose significant challenges to the structure determination process, which generally requires highly homogeneous proteins samples. In this book chapter, we provide a brief overview of protein disorder, describe various bioinformatics resources that have been developed in recent years for their characterization, and give a general outline of their applications in various types of structural genomics projects. Traditionally, disordered segments were filtered out to optimize the yield of structure determination pipelines. However, it is becoming increasingly clear that the structural characterization of proteins cannot be complete without the incorporation of intrinsically disordered regions.

Key words Intrinsically unstructured protein, Natively unfolded protein, Detection of protein disorder, Target selection and optimization, Conformational ensembles

1 What Is Protein Disorder?

According to the structure-function paradigm, the proper functioning of proteins requires a well-defined structure that ensures the precise orientations of atoms necessary for molecular recognition and catalysis. The general validity of this notion has been reinforced by a large number of successful structure-function studies, especially in the case of various enzymes, receptors, and structural proteins. This has also been the guiding principle of structural biology studies that have led to structure determination of close to 90,000 protein structures collected in the Protein Data Bank (PDB) [1]. For many decades, there were only sporadic examples of proteins that did not comply with this view. However, the advent of new experimental techniques combined with the wealth of protein sequence data provided by genome sequencing projects has showed that a large number of proteins do not adopt a well-defined structure under physiological conditions.

In a number of cases, detailed molecular characterization revealed that these proteins are functional, but their function relies on sampling a large set of alternative spatial (3D) conformations rather than a limited one as in the case of proteins that comply with the structure-function paradigm. This new group of proteins and protein regions has been termed intrinsically disordered (IDPs/IDRs) [2–4].

As we just mentioned, disordered proteins can be characterized by an ensemble of rapidly interconverting conformations. However, the detailed properties of this ensemble can vary greatly, making protein disorder a heterogeneous phenomenon [3, 5]. Indeed, using a combination of experimental techniques, various forms (flavors) of protein disorder can be discriminated [6, 7]. While some proteins are known to be fully disordered, many proteins are composed of both ordered and disordered regions of various lengths [8–10]. Disordered segments can correspond to loop regions, flexible *termini*, or linker regions between ordered domains [5]. Some disordered regions can exist in a form similar to random coils, while others are similar to molten globules exhibiting a compact but disordered state with some secondary structure content [6]. In essence, there is a continuum of structural states from fully disordered states to folded structures with various amounts of secondary/tertiary residual structure [5].

IDPs are also heterogeneous in terms of their function [11, 12]. In the case of IDPs functioning as entropic springs or spacers, the function directly originates from fluctuating among a large number of conformations and resisting changes that would reduce their conformational freedom [3]. As flexible linkers, disordered segments can also influence the distance and the orientation of adjacent ordered domains [5]. In the most typical scenario, however, disordered protein segments are involved in binding to other macromolecules (proteins, DNA, or RNA) [13]. During this process, they can undergo a disorder-to-order transition and become partially or fully ordered. Under such circumstances, they adopt a well-defined conformation that can be studied via traditional structure determination techniques. As part of a complex, they exhibit different structural features and can adopt regular secondary structure alongside coil and polyproline conformations. Among regular secondary structures, alpha helices are most common, but beta strands are also observed. In known examples, the length of the segment that is actually binding to the protein partner varies from just a few residues to up to 70 residues. The interface formed by disordered proteins has distinct characteristics compared to the interface between ordered proteins [14, 15]. Generally, some flexibility persists even in the complexed form of disordered proteins. This phenomenon, termed fuzziness, poses further challenges to the structure determination of complexes involving disordered proteins [16].

In general, protein disorder has several functional advantages including the adaptability in binding, high density of relatively compact functional sites that can act in a competitive or cooperative

manner, weak but specific binding, and frequent regulation by posttranslational modifications [17]. These properties make disordered proteins excellent candidates for the integration of various dynamic cellular signals [9] and explain their prevalence in regulatory and signaling processes [18, 19]. The presence of disordered regions is generally believed to be a quite common phenomenon particularly in eukaryotic proteins [19, 20]. Using two popular disorder prediction methods (IUPred and MD) to estimate the amount of disorder in complete proteomes, it was shown that 36–42 % of eukaryotic sequences contain at least one long (i.e., more than 30 residues) disordered segment, while the corresponding numbers for Archaea and Bacteria are only 7–13 %, depending on the method [21]. The biological importance of protein disorder is further underlined by the fact that many disordered proteins are associated with various diseases [22, 23]. Recognition of the biological importance and prevalence of protein disorder motivated the development of various bioinformatic resources for their description and characterization.

2 Databases of Ordered and Disordered Proteins

The Protein Data Bank is the largest existing repository of experimentally determined three-dimensional structures of proteins [1]. Currently (January 2014), this database holds almost 90,000 protein entries. Most of these structures were solved by X-ray crystallography and around 10 % by NMR. By taking into account sequence similarities, these protein structures can be split and clustered into several thousand different families [24]. While the PDB is primarily a collection of protein ordered regions (*see Note 1*), it also contains indirect information about protein disorder. In structures solved by X-ray crystallography, disordered segments are usually identified as regions that, although flanked by ordered protein segments, could not be assigned 3D coordinates experimentally [25]. These residues are said to be disordered, under the assumption that it is their conformational flexibility that makes their 3D arrangement difficult to describe using crystallographic techniques (*see Note 2*). Indication of protein disorder can also come from significant differences in residue-residue distances in different models of an NMR ensemble or in different chains found in the asymmetric unit of an X-ray structure [26]. It is important to remark, however, that residues with the above characteristics constitute less than 10 % of all residues in PDB proteins and occur in most cases in short segments at the proteins' *termini* [27, 28]. Some proteins deposited into the PDB only adopt a well-defined conformation as a result of interactions with cofactors, DNA, RNA, or other proteins. The PDB contains several hundred protein-protein complexes involving disordered proteins, which have been systematically analyzed [14, 15, 29].

Besides X-ray crystallography and NMR, protein disorder can be identified by a number of other experimental techniques, including circular dichroism spectroscopy (both far and near UV), protease sensitivity, and heat stability [30]. It is worth noting that there could be some disagreement between experimental methods with respect to the characterization of protein disorder, leading to inconsistencies in the assignment of protein disorder (*see Note 3*). The Database of Protein Disorder (DisProt) was the first database that aimed to collect and organize knowledge regarding the experimental characterization and the functional associations of IDPs [31, 32]. Disordered regions in DisProt are generally longer than segments annotated as disordered based on PDB structures with their length varying from 30 to over 18,000 residues. When this information is available, the database adds annotations about ordered segments for its entries, but for most proteins, annotation remains to date incomplete. Unfortunately, the size of the DisProt database is still very limited, as it currently holds annotation for less than 700 proteins. This represents only a small fraction of the available sequences that are expected to contain disordered regions. IDEAL, or the database of “Intrinsically Disordered proteins with Extensive Annotations and Literature,” also creates manual annotations for IDRs [33, 34]. This database collects experimental information from different sources including X-ray structures, NMR structures, circular dichroism studies, and other experiments. IDEAL takes special care in annotating IDRs that adopt 3D structures upon binding to their interaction partners. Also, to provide more complete annotation for its entries, IDEAL has recently added automatic annotations from DICHOT, a tool that combines automatic assignment of structural domains with IDR predictions [35]. The latest available release of IDEAL (IDEAL 30, August 2013) contains 340 annotated protein entries, with some overlap to DisProt [34].

To provide annotation for a larger number of proteins than available via DisProt or IDEAL, two databases have recently been established: MobiDB and D2P2. MobiDB provides annotation of disordered regions for most sequences in the UniProt database, combining experimental annotations from DisProt, X-ray structures, and NMR structures with results from several prediction methods [36]. For the final mark up of disordered vs. ordered regions, MobiDB uses a weighted consensus score that assigns higher weights to experimental annotations over predictions. D2P2, the database of Disordered Protein Prediction, runs a battery of disorder predictors on all protein sequences from a list of complete proteomes (totaling 1,765 as of January 2014) [37]. D2P2 integrates disorder prediction results with the (mostly structured) SCOP domains annotated using profile HMMs found in the SUPERFAMILY database [38].

With their slightly different focus, all these databases contain important information about the occurrence of disordered proteins and their functions (Table 1).

Table 1
List of databases for disordered proteins

Name	Web URL	Focus	Number of proteins
DisProt [31, 32] <i>Database of Protein Disorder</i>	http://www.disprot.org/	Manually curated database of proteins that lack fixed 3D structure in their putatively native states verified by experimental methods	694 proteins (1,529 disordered regions)
IDEAL [33, 34] <i>Intrinsically Disordered proteins with Extensive Annotations and Literature</i>	http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/	Collection of experimentally verified intrinsically disordered proteins (IDPs) with manually curated annotations on IDPs in locations, structures, and functional sites such as protein- binding regions and posttranslational modification sites together with references and structural domain assignments	340 proteins
MobiDB [36] <i>Database of protein disorder and mobility annotations</i>	http://mobidb.bio.unipd.it/	Centralized source for data on different flavors of disorder in protein structures. In addition to the DisProt and PDB X-ray structures, it includes experimental information from NMR structures and five different flavors of two disorder predictors (ESpritz and IUPred)	45,323,666 annotated proteins
D2P2 [37] <i>Database of Disordered Protein Predictions</i>	http://d2p2.pro/	A community resource for precomputed disorder predictions on a large library of proteins from completely sequenced genomes	10,429,761 sequences in 1,765 genomes from 1,256 distinct species
PE-DB [91] <i>Protein Ensemble Database</i>	http://pdb.vib.be/	A database for the deposition of structural information on IDP- and denatured protein ensembles based on NMR and SAXS data	5 proteins: 3,973 conformations of 40 ensembles in 11 entries

The list of various databases containing disordered proteins with their names, web address, main focus, and the number of proteins they contain

3 Prediction of Protein Disorder

As we have seen, experimental information on disordered proteins is still sparse. This underlines the importance of developing bioinformatics methods for their characterization. The first computational approach devoted specifically to this problem was developed in 1998 by the group of Keith Dunker [39]. Since then, over 50 methods have been published in the literature, and many of them are available as a web server or a stand-alone program [40, 41]. Currently, there are two main applications for disordered prediction methods. On the one hand, these methods can help in understanding the functional and evolutionary properties of disordered proteins and their association with various diseases both in small- and large-scale studies. On the other hand, disorder prediction methods have become integral part of protein target selection and construct optimization in structure determination efforts. Indeed, as IDPs represent a significant challenge to protein crystallization attempts, the capability to identify (and exclude from experimental characterization) disordered regions has been extremely relevant to the success of structural genomics projects. These projects pursue large-scale protein structure determination via high-throughput semiautomatic experimental (e.g., protein production) pipelines. In the next sections, we provide an overview of available bioinformatics resources for protein disorder, of their application in structural genomics projects, and we discuss how extended characterization of protein disorder will likely require adopting a new approach to protein structural studies.

3.1 Overview of Disorder Prediction Methods

Analysis of the sequences of experimentally verified disordered proteins has shown that IDPs have distinct sequence properties compared to globular proteins. In general, IDP sequences are enriched in polar and charged residues and depleted in bulky aliphatic and aromatic amino acids. Based on this observation, groups of disorder-promoting (KEPSQRA) and order-promoting amino acid residues (WCFIYVL) have been established [42]. Often disordered regions have biased amino acid composition characterized by an overwhelming presence of disorder-promoting amino acids and hence may overlap with low-complexity segments [3]. Several amino acid features were suggested to be related to protein disorder, including flexibility, aromatic content, secondary structure preferences, high net charge, and various scales related to hydrophobicity [43]. A single amino acid propensity scale forms the basis of the GlobPlot and FoldUnfold prediction methods [44, 45] (see Table 2 for a list of methods). Simple physical principles can also guide the prediction of disorder. It was suggested that high net charge and low average hydropathy is characteristic of disordered proteins [6]. This principle was implemented as a position-specific

Table 2

List of publicly available disorder prediction methods

Method	Web URL	Approach	Output	Speed
IUPred [47]	http://iupred.enzim.hu	Estimation of pairwise interaction energies; trained on globular proteins only	Graphical, text output, download	Fast
FOLDINDEX [46]	http://bip.weizmann.ac.il/fldbin/findex	Amino acid propensities; uses the combination of net charge and hydrophobicity calculated with a sliding window	Graphical, text output, download	Fast
GlobPlot [45]	http://globplot.embl.de/	Amino acid propensity; based on propensity to be in regular secondary structure as opposed to be in coil	Graphical, text output, download	Fast
DisEMBL [54]	http://dis.embl.de/	Neural network; three separate predictions, for residues in loops, for residues in loops with high B-factor, and for residues in REMARK 465	Graphical, text output, download	Fast
PONDR VL2 [7]	http://www.dabi.temple.edu/disprot/predictor.php	Neural network; predictions are based on various sequence features (18 amino acid frequencies, average flexibility, and sequence complexity) trained on long disordered segments to recognize various flavors of disorder	Text output	Fast
PONDR VL3 [51]	http://www.dabi.temple.edu/disprot/predictor.php	Ensemble of Neural Networks; trained using 20 sequence attributes on long disordered segments	Text output	Fast (without PsiBlast)
PONDR VSL2 [49]	http://www.dabi.temple.edu/disprot/predictor.php	Support vector machines (SVM); specific predictors for long and short disordered segments combined using an additional predictor	Text output	Fast (without PsiBlast)

(continued)

Table 2
(continued)

Method	Web URL	Approach	Output	Speed
RONN [55]	https://app.strubi.ox.ac.uk/RONN/	Biobasis function neural network; similarity to known disordered segments	Graphical, text output	Fast
DISOPRED2 [116]	http://bioinf.cs.ucl.ac.uk/psipred/?disopred=1	SVM; trained on missing X-ray residues, very low false-positive rate	Graphical, text output, download	Slow
OnD-CRF [61]	http://babel.ucmp.umu.se/ond-crf/	Conditional random fields; predicting the transition between structured and mobile or disordered regions in proteins, using features generated from the amino acids sequence and from secondary structure prediction	Email, graphical text	Slow
UCON [52]	https://www.predictprotein.org/	Neural network; contact-based prediction of disordered sites	Text output	Cached
NORsp [53]	https://www.predictprotein.org/	Neural network; long regions without predicted secondary structure elements	Text output	Cached
CSpritz [60]	http://protein.bio.unipd.it/cspritz/	Bidirectional recursive neural networks (BRNN); combination of machine-learning and modeling techniques trained on PDB structures (X-ray and NMR) and DisProt data	Text output	Slow
ESpritz [59]	http://biocomp.bio.unipd.it/espritz/	Bidirectional recursive neural networks and trained on three different flavors of disorder, faster than CSpritz	Text output	Fast
PONDR-FIT [67]	http://www.disprot.org/pondr-fit.php	Meta server; neural network trained on fully ordered and disordered proteins	Graphical, text output, download	Fast

Poodle L, S, W [50, 56, 117]	http://mbs.cbrc.jp/poodle/poodle.html	SVMs and other machine-learning methods; separate predictions for short disorder regions prediction, long disorder regions prediction, and unfolded protein prediction	Graphical, text output	Cached
PreDisorder [58]	http://casp.rnet.missouri.edu/predisorder.html	1D recursive neural network using PSI-Blast profiles and predicted secondary structures and solvent accessibility	Email response, download	Slow
DISOclust [57]	http://www.reading.ac.uk/bioinf/DISOclust/	Predictions are based on heterogeneity of 3D structural models combined with Disopred2 predictions	Email response, download	Slow
Genesilico-Metadisorder [65]	http://genesilico.pl/metadisorder/	Meta server; trained on pdbRemark465, CASP7, and Disprot dataset, incorporates fold recognition method	Graphical, text output	Cached
MD [68]	https://www.predictprotein.org/	Meta server; neural network based meta-predictor combining orthogonal approaches	Graphical, text output	Cached
DisMeta [64]	http://www-nmr.cabm.rutgers.edu/bioinformatics/disorder/	Meta server; consensus prediction	Email response	Slow
metaPrDOS	http://prdos.hgc.jp/cgi-bin/meta/top.cgi	Meta server; combined with template-based modeling	Email response	Slow
MFDp [66]	http://biomine-ws.ecc.uialberta.ca/MFDp.html	Meta server (multilayered fusion-based disorder predictor); ensemble of SVMs specialized for the prediction of short, long, and generic disordered regions	Email response	Slow

The list of publicly available disorder prediction methods, giving for each method the name, the web address, a short description of the underlying approach, the type of the output, and the speed of server response. Some methods cache earlier predictions; therefore, predictions for cached proteins are returned quickly

prediction method in the FoldIndex server [46]. The IUPred method predicts protein disorder by estimating the pairwise energy content encoded by a protein sequence [47]. The basic idea of this method is that regions that cannot form enough favorable interactions within their sequence environment remain disordered [48].

The prediction of protein disorder can be framed as a classic binary classification problem and targeted with various machine-learning methods. Many disordered prediction methods are based on such approaches. The most commonly used techniques are support vector machines [19, 49, 50] and neural networks [51–54], but other approaches such as biobasis function neural networks [55], or various clustering techniques [56, 57], are also utilized. Methods like recursive neural networks [58–60] or conditional random fields [61] can achieve improved performance by capturing the interdependence in the disorder tendency of sequence-neighboring residues. One of the main advantages of machine-learning methods is that they can easily incorporate additional information beside the amino acid sequence. These are, for example, evolutionary profiles in the form of position-specific scoring matrices, predicted secondary structure, solvent accessibility, and flexibility. Predictions can also be enhanced using information obtained from template-based modeling [62]. However, while machine-learning methods are relatively straightforward to build, they give very little insight on the underlying causes of protein disorder.

When building and using disorder prediction methods, one has to be aware of the various flavors of protein disorder. For example, disordered regions collected from the PDB and from DisProt differ not only in terms of their length but also in terms of their amino acid composition [3]. Ignoring these differences can limit the performance of a prediction method, as methods trained on one type of disorder perform less well on the other type. PONDR VSL2 was the first method to take this into account [49, 63]. It uses a combination of support vector machine predictors for both short and long disordered regions with the final prediction being a weighted average determined by a second layer predictor. More recent methods are overwhelmingly all meta-servers. These methods aim to achieve improved predictions by combining the output of individual methods that can be specific to certain flavors of disorder [64–68].

3.2 Performance of Disorder Prediction Methods

Since 2004, the Critical Assessment of protein Structure Prediction (CASP), a biannual, community-wide blind experiment launched in 1994 (<http://predictioncenter.org/>), has had a section devoted to the assessment of disorder predictors [25]. The evaluation is based on PDB structures that are only released after predictions are made, thus ensuring that submitted predictions are “blind” with respect to experimental structural information. In the most recent editions, the number of target structures released for assessment has been around one hundred, with ordered residues in target

structures usually largely outnumbering disordered ones. For example, in the last round of CASP (CASP10), 1,664 residues in the assessed targets were identified as disordered, constituting 6.8 % of all residues [28]. The most commonly used measures for evaluation are Matthews coefficient (MCC), balanced accuracy (ACC), and area under ROC curve (AUC) [28, 69]. These measures usually ensure a balanced assessment of sensitivity and specificity. In the CASP10 dataset, best performing groups can achieve AUC above 0.9, MCC score above 0.5, and ACC score around 0.75. However, a pronounced decrease in performance was observed when considering only internal positions or longer disordered segments. The best predictors in CASP10 were only moderately more accurate than best methods in previous rounds, indicating a modest progress in recent years [28].

As most CASP targets are solved by X-ray crystallography, they typically contain relatively short disordered regions, which are not representative of the type of disorder observed in functionally relevant proteins (typically longer disordered regions) [28]. Recently, 16 different publicly available methods were tested using a benchmark dataset that contained disordered segments of at least 30 residues collected from both PDB and DisProt entries [70]. When using the AUC measure, the best performing methods were MFDp, PONDR-FIT, MD, and VSL2. However, when using different evaluation criteria, other methods emerged as top predictors. These results suggested that there is no universally superior predictor and that some of the top-performing methods are complementary [70]. Largely due to the limited collection of experimentally characterized disordered regions, obtaining a good unbiased estimate of the performance of disorder prediction methods remains a challenging task.

Despite differences among prediction methods, conclusions derived from large-scale studies show no dependence on the choice of prediction methods (*see, e.g., refs. 71–73*). Predictions generated by different approaches often show a clear consensus, with disagreement between the prediction outputs often corresponding to “dual personality” or borderline cases. Figure 1 shows the case of calcineurin, featuring both instances in which the output of disorder prediction methods is in agreement and instances in which the methods instead disagree. A typical situation in which methods may disagree involves disordered regions that fold upon binding. These regions are disordered in isolation but at the same time feature a pronounced tendency for being ordered. This dual personality manifests itself when the proteins find a binding partner that induces a disorder-to-order transition. Depending on the computational tool, such regions can be predicted as completely disordered, completely ordered, or as borderline cases [41]. Additional examples of disordered regions featuring a significant tendency to be ordered include coiled coils and molten globule-like disordered segments [41]. The complexity of these cases suggests that the

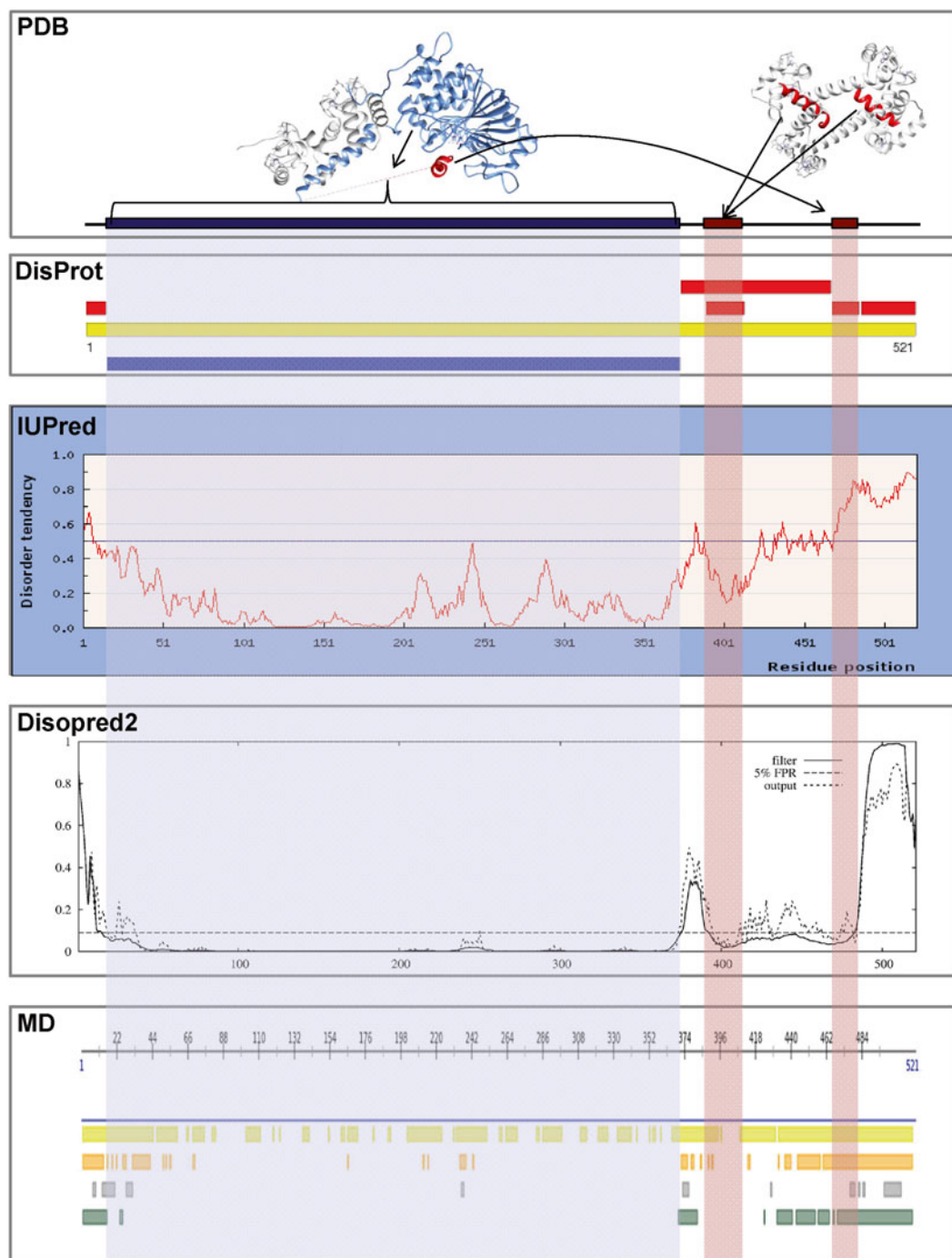


Fig. 1 Various types of disorder in human calcineurin. Calcineurin is a calcium-dependent serine-threonine phosphatase, and it is composed of a catalytic A subunit (*shown here*) and a regulatory B subunit [40]. The plot shows the regions that can be found in the PDB and annotations according to DisProt database [32], as well as the output from three disorder prediction methods, IUPred [47], Disopred2 [116], and MD [68]. MD (*green, shown at the bottom*) is a meta-server that is based on the prediction from PROFbval (*yellow*), UCON (*orange*), and Norsnet (*gray*) besides IUPred and Disopred2. This protein contains several disordered regions, including

strict categorization of protein regions into ordered and disordered is a great oversimplification and that fully embracing this complexity may lead to improved prediction methods.

3.3 Application of Protein Disorder Prediction Methods in Structural Genomics

Structural genomics initiatives employ high-throughput protein production and structure determination pipelines to solve protein three-dimensional structures [74]. One of the main goals of structural genomics projects over the years has been to try to provide structural information for most of the known protein families (especially those that are taxonomically diverse). In order to achieve this, one or more structural representatives within a family are solved experimentally, while the rest is leveraged using comparative modeling techniques. To make this process more efficient and help abating costs, target selection strategies have often focused on identifying, and then excluding, so-called high-hanging fruits. These are sets of families or proteins that present considerable challenges to structural determination either in general or under the specific experimental protocols implemented in a given pipeline. Membrane, coiled coil, and disordered proteins are all examples of such high-hanging fruits.

The presence of disordered segments is problematic at each step of a structure determination pipeline that follows cloning [75, 76]. For example, when heterologously expressed, the many IDPs known to be involved in signaling and regulation might be toxic for the expression host. Also, due to their low conformational stability, IDPs are more prone to proteolytic degradation [77]. Although in general protein disorder increases solubility due to high polar content, in certain cases, short patches of exposed hydrophobic residues located within disordered segments can promote aggregation [78]. Finally, disordered proteins and protein regions are problematic for both X-ray crystallography and NMR structure determination protocols, as these techniques rely on homogeneous conformational ensembles to amplify experimental signals [79]. Indeed, the observed abundance and length distribution

Fig. 1 (continued) the first 13 amino acids that were missing from the electron density map, and multiple fragments of the C-terminal region that were shown to be disordered by various experimental methods, according to the DisProt annotations. The C-terminal region harbors the autoinhibitory peptide that binds to the active site when the enzyme activity is turned off (shown here on the *left*, PDB code 1AUI [118]), but increased calcium level leads to the release of the peptide and the activation of the enzyme. Calcineurin also has a calmodulin-binding site located within the disordered region that becomes ordered upon binding (PDB code 2R28, on the *right*). The region corresponding to the ordered catalytic domain is *shaded light gray* and the two binding regions located within disordered segments are *shaded light red*. For most regions, including the ordered catalytic domain and disordered regions in the N- and the very C-terminal regions, there is a good agreement among the methods. There is no consistency, however, for the autoinhibitory peptide and the preceding disordered region that is predicted as ordered by Disopred2, disordered by MD, and as a borderline case by IUPred. In contrast, the calmodulin-binding site is predicted ordered by all three methods

of missing density regions in the PDB indicates that crystal structures can tolerate only a small extent of disorder, and when present, such regions tend to be limited to relatively short segments [80]. Because of these reasons, in recent years, disordered prediction methods have become important tools for protein construct design and for sample preparation optimization in both NMR and crystallization studies [64, 76, 79, 81].

4 Disordered Regions: From Scourge of Structural Studies to Targets?

The size of the known protein sequence universe is rapidly expanding, largely due to genome sequencing projects. In each newly sequenced genome, the majority of the predicted proteins belong to already established protein families. However, there is no sign of saturation of sequence space: each new sequenced genome adds many novel sequences and families. The number of known protein structures deposited into the PDB database has also been growing rapidly. However, recent estimates suggest that only a small fraction of new protein structures solved today represent a previously structurally uncharacterized protein family [24]. Despite enormous progress in protein production, crystallization, and structure determination methodologies and technologies, nearly half of the content of the SwissProt database is still without structural representatives [82]. Predictably, the set of proteins and families with no structural representative is enriched with transmembrane and intrinsically disordered regions [24]. More systematic efforts and novel approaches will be required to achieve good structural coverage of these classes of proteins.

Among proteins that are recalcitrant to structure determination, membrane proteins represent an interesting case study. Traditionally divided into alpha-bundle and beta-barrel proteins, membrane proteins constitute an important fraction of the proteome of living organisms [83]. Among their many functions, these proteins are heavily involved in signaling and transport. Because they mediate communication between the cell and the external environment, they are of great interest to biomedical research (e.g., ion channels, G-protein-coupled receptors). Unfortunately, membrane proteins (especially those belonging to the alpha-bundle type) present structural biologists with several challenges. They have typically low native expression levels, and attempts to overexpress them are often toxic to the expression host. Purification often proves difficult as it generally requires extracting the proteins from their native membrane environment with the use of detergents. This explains why, as we mentioned above, membrane proteins are greatly underrepresented in structural databases, constituting less than 2 % of all proteins in the PDB [84]. In fact, membrane proteins were for quite some time

regarded as the high-hanging fruits “par excellence” by structural genomics consortia and carefully filtered out of these initiatives’ target lists. Membrane proteins’ abundance and multifarious functional roles, however, have ensured continued efforts to improve our understanding of the techniques required to express, purify, and stabilize these proteins. Indeed, starting in 2005, structural genomics initiatives were established that specifically targeted membrane proteins [85, 86]. Although progress has been relatively slow and much work remains to be done [84, 87], membrane proteins are now recognized as challenging but increasingly solvable targets, showing that researchers’ dogged single-mindedness can tame even the most problematic proteins. Today, it could be argued that the baton of “most challenging targets” has been passed from membrane proteins to disordered proteins. Similar to what has been and is being done for membrane proteins, we now need to step up efforts to expand functional and structural characterization of disordered protein regions.

It may seem odd to talk about structural characterization of IDPs; however, the ensemble of conformations adopted by intrinsically disordered proteins is never completely devoid of structural elements, as different conformations have different energies and are consequently adopted with different probabilities. While knowledge of the details of these dynamic ensembles can be very useful to understand how function arises from the disordered state, their experimental characterization remains a very challenging task and generally requires the combination of different methodologies [88]. Various types of NMR experiments complemented with SAXS measurements can indicate the presence of transient secondary structure elements or long-range tertiary contacts [89]. Using the experimental data as constraints, computational techniques can be used to derive a restricted subset of conformations that are in agreement with observed hydrodynamic behavior and local structural preferences [90]. However, one has to be aware that these ensembles are a neither precise nor complete representation of disordered states, but rather models that fit a specifically defined subset of data. These approaches cannot provide unique solutions given the extreme conformational freedom of IDPs and the limited amount of data that is available. However, they can help to characterize disordered proteins’ conformational ensembles capturing some elements of the fleeting structural features of these proteins. The pE-DB database was recently launched to aid the development and consolidation of standards for describing the structural ensembles of intrinsically disordered and unfolded proteins [91]. This database currently holds only 11 entries, including the conformational ensembles for human Tau protein (Table 1). Alongside with the emerging field of IDPs, it is increasingly recognized that the description of biomolecules as static structures is inadequate and proteins must be described as ensembles of thermally accessible conformers. Therefore, the new approach to

structural characterization embraced by pE-DB can have a significant impact on globular proteins studies as well, possibly marking a paradigm shift in structural biology.

5 Disorder Prediction Programs

5.1 Technical Aspects of Disorder Prediction Methods

Although more than 50 methods have been published in the literature [40], less than half of them are publicly available either as a web server or as a downloadable program (Table 2). Generally, predictions are carried out for one sequence at a time, although multiple sequences can also be submitted in the case of a few methods. The input of disorder prediction methods is then usually a single amino acid sequence, either in FASTA format or in a simple format without the header. It should be noted that there could be length limits on the submitted sequences (*see* **Note 4**). Certain methods can also take a UniProtKB ID or accession number as an input.

Web servers can present the output in either text or graphical format, while methods that require longer time to carry out calculations often return the results via email. In the most common output format, the disordered tendency is characterized in a position-specific manner. A score between 0 and 1 is assigned to each residue in the amino acid sequence. Positions with a score above 0.5 are regarded as disordered, while those with a score below 0.5 as ordered. Some deviations from the standard format also exist. For example, the DISOPRED method uses a cutoff different than 0.5, while the FoldIndex method's [46] prediction score is not restricted to the [0,1] range. Differences among predictors can also emerge as a consequence of the different weights that they place on the mis-prediction of ordered versus disordered residues during method parameterization. This property can be characterized by the false-positive rate, which determines the amount of ordered residues predicted as disordered. Indeed the score cutoff of DISOPRED2, discussed above, reflects the desire to keep the false-positive rate to very low levels (estimated at 2 %) [19]. On the other hand, the false-positive rate of the PONDR methods is usually quite high, resulting in more regions predicted as disordered at the cost of lower specificity.

The position-specific profiles that reflect disorder propensity that are returned by predictors are often quite noisy. For this reason, raw prediction outputs are generally subjected to some type of smoothing, using either a sliding window technique or a second layer of prediction. Even from smoothed position-specific scores, however, it is often not straightforward to identify which regions correspond to ordered domains. To address this problem, the GlobPlot and IUPred methods provide domain-level predictions [45, 47]. Predictions at the full protein level also exist, including

the charge-hydrophobicity plot and the cumulative distribution function (CDF), which distinguish ordered and disordered proteins based on the distribution of prediction scores [92, 93].

Generating disorder predictions for a single protein can take just a few seconds or several hours, depending on the method. The speed of a method can be an important factor especially in the case of large-scale analyses. Methods based on the physicochemical properties of disorder, like IUPred, GlobPlot, or FoldIndex, are generally fast. In contrast, methods that require PSI-BLAST profiles [94], or additional information that rely on such profiles, like secondary structure or solvent accessibility predictions, usually take longer. It should also be noted that while considering information from methods that predict additional protein structural features can boost performance, the effect on disorder predictions is usually smaller when compared to other structure prediction problems. One likely explanation for this is that many methods have been developed for globular proteins only and their prediction accuracy may be lower when they are applied to disordered regions (*see Note 5*). Meta-servers generally include a mixture of fast and slow methods; as a consequence, they also operate on the slower timescale.

5.2 Using Disordered Prediction Methods in Construct Optimization and Design

The first step in this process should be carrying out disordered predictions, preferably with multiple methods. Depending on the target dataset, it might also be possible to directly use predictions found in MobiDB or D2P2. An example of MobiDB consensus disorder assignment making use of several disorder prediction methods and of information from the PDB is given in Fig. 2. It is highly recommended that regions of at least 30 residues predicted as disordered be removed from the construct [76]. If the percentage of disordered residues remains above 30 %, further construct optimization may be needed. Finally, the lack of mostly ordered regions of at least 30 residues suggests that the protein is unlikely to be able to form crystals suitable for structure determination [76]. If predictions show a clear consensus, results' interpretation is straightforward. If not, information about protein disorder can be complemented by a variety of methods. Low-complexity region prediction, using, for example, the SEG algorithm [95], can help in confirming the presence of disordered regions. In contrast, predicted signal sequences and transmembrane regions can indicate regions that potentially cause problems in the structure determination process even when they are predicted as ordered by disorder prediction methods, as it commonly happens [64]. The existence of homologues for the target protein with experimentally verified ordered structures or, vice versa, experimentally known to be disordered can also help interpreting the prediction results. However, collecting evolutionary information about disordered

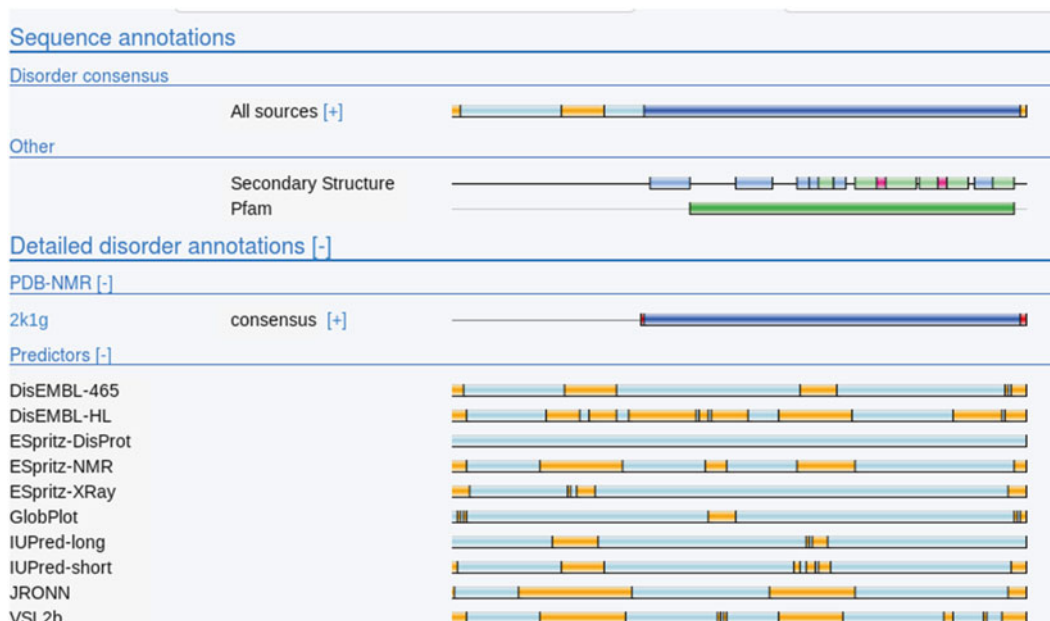


Fig. 2 Example output from the MobiDB database [36]. The *plot* shows the output for the spr lipoprotein (murein DD-endopeptidase MepS/Murein LD-carboxypeptidase) from *Escherichia coli*. The *MobiDB annotations* include various predictions for the given sequence, the structural annotations, and the consensus prediction. This protein was an NESG target (ER541), and the structure of this protein was determined as a result of a successful construct optimization based on consensus disorder prediction methods, after removing the N-terminal part [64]

protein segments can be more difficult due to their distinct evolutionary properties (*see Note 6*).

The lack of agreement between disorder prediction methods is common in regions that are disordered in isolation but can undergo a disorder-to-order transition upon binding to other molecules [41]. Such regions often overlap with linear motifs [96]. These specific regions involved in protein interactions can also be predicted from the amino acid sequence using a variety of methods. Current methods rely on simple biophysical models [97], sequence properties [98, 99], recognition of specific sequence motif [100], or on patterns of evolutionary conservation [101, 102]. Disordered regions can also be involved in binding of DNA and RNA, usually indicated by the abundance of positively charged residues. Posttranslational modifications can also influence the disorder status of proteins (*see Note 7*).

Multidomain proteins often misfold in prokaryotic systems and also often exceed the size limitation of high-throughput NMR structure determination techniques [64]. Domain parsing can help to circumvent these issues. In some cases, disorder prediction methods alone can be used to predict domain boundaries, especially when a longer disordered segment separates two ordered domains.

However, a short predicted disordered region might not be a linker between structural domains but rather a flexible loop located within a single domain. To be able to discriminate between these two cases, one can rely on domain boundary prediction methods [103, 104]. Multiple sequence alignments of homologous proteins can also help to identify possible structural domain boundaries. Various resources offer information about homologous domains, including InterPro [105], Pfam [106], and SMART [107].

In many cases, a more complex strategy might be needed to obtain constructs that express at high levels, remain soluble, and are generally suitable for structural analysis. High-throughput cloning and expression platforms allow for efficient production and processing of several alternative constructs, for example, by varying the *termini* of a targeted domain [64, 79]. These constructs can be further optimized using small-scale expression and solubility screening. The suitability of the constructs for structure determination can be checked by various experimental methods, including HDQC NMR, ¹⁵N nuclear relaxation rate, or hydrogen-deuterium mass spectroscopy data [108, 109].

6 Notes

1. The common assumption is that proteins in the PDB are ordered. However, some proteins deposited into the PDB only adopt a well-defined 3D conformation as a result of crystal contacts, interactions with cofactors, DNA, RNA, or other proteins. In order to create a more reliable database of ordered proteins, PDB structures containing DNA, RNA molecules, and cofactors are often filtered out.
2. Information about disordered regions in the PDB can be collected by comparing the ATOM and SEQRES records in the PDB entry's flat files. Although this information can also be available in the REMARK 465 lines, the corresponding annotations are often incomplete. It is worth noting that the lack of spatial coordinates in a PDB entry can arise from causes other than intrinsic disorder. For example, multidomain proteins connected by flexible hinges can cause one of the domains to adopt a heterogeneous or dynamic orientation, resulting in missing density in the solved crystal structure [76]. Certain regions could also lack spatial coordinates as a result of proteolysis events that are often not annotated properly, or other annotation errors.
3. There is still some uncertainty on what constitutes a disordered segment. There are two main reasons for this. First, protein disorder has been shown to be a heterogeneous phenomenon. Second, the assignment of a protein segment to the disordered

category may depend on the experimental methods and on the conditions under which it is being studied. For example, a comparison of structures of identical proteins solved in different conditions indicated hundreds of fragments found as ordered in some structures and as disordered in others [110]. Similar differences can also be observed when considering other experimental techniques. This can lead to inconsistencies in databases regarding order and disorder assignments and can limit the performance of prediction methods that are developed based on such data.

4. Some disorder prediction methods set a minimum and/or maximum limit on the length of the sequences that can be submitted for prediction. The maximum limit, usually 1,000 residues, is often set for practical reasons. When in need of submitting longer sequences, the input can be chopped into smaller parts. However, while in most prediction methods distant sequence elements do not influence greatly local disorder predictions, it should be noted that the sequence neighborhood affecting the prediction can range from 10 to 20 residues (e.g., [19]) to up to 100 (e.g., [48]).
5. Secondary structure prediction methods have been developed using globular protein structures, and they should hence be used with caution in disordered regions. It was suggested that longer regions (>70) without any predicted secondary structure can correspond to disordered segments [53]. However, the reverse is not true as predicted secondary structure elements within disordered segments can correspond to preformed structural elements or to the structure adopted in a complexed form [111]. From the perspective of the proteins that host such regions, these regions should be still regarded as disordered, at least when the protein hosting them is considered in isolation. In most cases, however, secondary structure predictions can help in identifying the boundaries of ordered regions. Also, cutting into a predicted secondary structure element should be avoided in construct design.
6. Structured domains are often evolutionary conserved units, while disordered regions are in general less conserved [112]. There are, however, exceptions from these general trends. One remarkable example is the human nonhistone chromosomal protein HMG-17. This protein is highly disordered and contains low-complexity sequences. However, it is highly conserved and part of a Pfam family (HMG14_17). This should be kept in mind when evolutionary conserved regions are used. For example, the Pfam database [106] discriminates domains, families, repeats, and motifs, all of which correspond to evolutionary conserved sequence elements. From these categories, however, only the presence of domains and repeats (when

present in more than one copy) should be considered as likely structured.

7. Currently, prediction methods cannot handle nonstandard amino acids, and their presence in sequences submitted for prediction is either ignored or cause an error. Posttranslational modifications, however, can significantly change the behavior of disordered regions. One of the most commonly occurring PTM is phosphorylation, which is often observed switching on or off binding events involving disordered regions [113]. PTMs can also induce a transition to a more extended or compact structure [114], or even transform an ordered structure to a completely disordered one [115]. From the prediction point of view, the effect of PTMs can sometimes be modeled by carrying out the prediction after replacing residues undergoing modifications with the chemically most similar standard amino acid.

Acknowledgements

This work was supported by grants of the Hungarian Scientific Research Fund (OTKA) K108798 and NK100482 [Z. D. and I. S.] and Wellcome Trust WT077044/Z/05/Z [M.P.]

References

1. Berman HM et al (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242
2. Tompa P (2002) Intrinsically unstructured proteins. *Trends Biochem Sci* 27:527–533
3. Dunker AK et al (2001) Intrinsically disordered protein. *J Mol Graph Model* 19:26–59
4. Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 293:321–331
5. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6:197–208
6. Uversky VN, Gillespie JR, Fink AL (2000) Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins* 41:415–427
7. Vucetic S et al (2003) Flavors of protein disorder. *Proteins* 52:573–584
8. Pentony MM, Jones DT (2010) Modularity of intrinsic disorder in the human proteome. *Proteins* 78:212–221
9. Gibson TJ (2009) Cell regulation: determined to signal discrete cooperation. *Trends Biochem Sci* 4:471–482
10. Bhattacharyya RP et al (2006) Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. *Annu Rev Biochem* 75:655–680
11. Tompa P (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett* 579:3346–3354
12. Xie H et al (2007) Functional anthology of intrinsic disorder. I. Biological processes and functions of proteins with long disordered regions. *J Proteome Res* 6:1882–1898
13. Dyson HJ, Wright PE (2002) Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol* 12:54–60
14. Meszaros B et al (2007) Molecular principles of the interactions of disordered proteins. *J Mol Biol* 372:549–561
15. Vacic V et al (2007) Characterization of molecular recognition features, MoRFs, and their binding partners. *J Proteome Res* 6:2351–2366
16. Fuxreiter M, Tompa P (2012) Fuzzy complexes: a more stochastic view of protein function. *Adv Exp Med Biol* 25:1–14

17. Tompa P (2012) Intrinsically disordered proteins: a 10-year recap. *Trends Biochem Sci* 37:509–516
18. Iakoucheva LM et al (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* 323:573–584
19. Ward JJ et al (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 337:635–645
20. Dunker AK et al (2000) Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* 11:161–171
21. Schlessinger A et al (2011) Protein disorder—a breakthrough invention of evolution? *Curr Opin Struct Biol* 21:412–418
22. Uversky VN, Oldfield CJ, Dunker AK (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* 37:215–246
23. Uversky VN et al (2009) Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. *BMC Genomics* 10 Suppl 1:S7
24. Mistry J et al (2013) An estimated 5 % of new protein structures solved today represent a new Pfam family. *Acta Crystallogr D Biol Crystallogr* 69:2186–2193
25. Melamud E, Moult J (2003) Evaluation of disorder predictions in CASP5. *Proteins* 53 Suppl 6:561–565
26. Bordoli LF, Kiefer F, Schwede T (2001) Assessment of disorder predictions in CASP7. *Proteins* 69 Suppl 8:129–136
27. Le Gall T et al (2007) Intrinsic disorder in the Protein Data Bank. *J Biomol Struct Dyn* 24:325–342
28. Monastyrskyy B et al (2014) Assessment of protein disorder region predictions in CASP10. *Proteins* 82 Suppl 2:127–137
29. Gunasekaran K, Tsai CJ, Nussinov R (2004) Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J Mol Biol* 341:1327–1341
30. Eliezer D (2009) Biophysical characterization of intrinsically disordered proteins. *Curr Opin Struct Biol* 19:23–30
31. Vucetic S et al (2005) DisProt: a database of protein disorder. *Bioinformatics* 21:137–140
32. Sickmeier M et al (2007) DisProt: the Database of Disordered Proteins. *Nucleic Acids Res* 5(Database issue):D786–D793
33. Fukuchi S et al (2012) IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature. *Nucleic Acids Res* 40(Database issue):D507–D511
34. Fukuchi S et al (2014) IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic Acids Res* 42: D320–D325
35. Fukuchi S et al (2009) Development of an accurate classification system of proteins into structured and unstructured regions that uncovers novel structural domains: its application to human transcription factors. *BMC Struct Biol* 9:26
36. Di Domenico T et al (2012) MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics* 28: 2080–2081
37. Oates ME et al (2013) D(2)P(2): database of disordered protein predictions. *Nucleic Acids Res* 41(Database issue):D508–D516
38. Wilson D et al (2009) SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res* 37(Database issue):D380–D386
39. Li X et al (1990) Predicting protein disorder for N-, C-, and internal regions. *Genome Inform Ser Workshop Genome Inform* 10: 30–40
40. He B et al (2009) Predicting intrinsic disorder in proteins: an overview. *Cell Res* 19: 929–949
41. Dosztanyi Z, Meszaros B, Simon I (2010) Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Brief Bioinform* 11:225–243
42. Williams RM et al (2001) The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pac Symp Biocomput*:89–100
43. Xie Q et al (1998) The sequence attribute method for determining relationships between sequence and protein disorder. *Genome Inform Ser Workshop Genome Inform* 9:193–200
44. Galzitskaya OV, Garbuzynskiy SO, Lobanov MY (2006) FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics* 22:2948–2949
45. Linding R et al (2003) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res* 31:3701–3807
46. Prilusky J et al (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* 21:3435–3458

47. Dosztanyi Z et al (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21:3433–3434
48. Dosztanyi Z et al (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* 347:827–839
49. Obradovic Z et al (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* 61 Suppl 7:176–182
50. Hirose S et al (2007) POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics* 23:2046–2053
51. Predicting intrinsic disorder from amino acid sequence. *Proteins* 53 Suppl 6:566–572
52. Schlessinger A, Punta M, Rost B (2007) Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics* 23:2376–2384
53. Liu J, Rost B (2003) NORSp: predictions of long regions without regular secondary structure. *Nucleic Acids Res* 31:3833–3835
54. Linding R et al (2003) Protein disorder prediction: implications for structural proteomics. *Structure* 11:1453–1459
55. Yang ZR et al (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 21:3369–3376
56. Shimizu K et al (2007) Predicting mostly disordered proteins by using structure-unknown protein data. *BMC Bioinformatics* 8:78
57. McGuffin LJ (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics* 24:1798–1804
58. Deng X, Eickholt J, Cheng PJ (2009) PreDisorder: ab initio sequence-based prediction of protein disordered regions. *BMC Bioinformatics* 10:436
59. Walsh I et al (2012) ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* 28:503–509
60. Walsh I et al (2011) CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. *Nucleic Acids Res* 39(Web Server issue):W190–W196
61. Wang L, Sauer UH (2008) OnD-CRF: predicting order and disorder in proteins using conditional random fields. *Bioinformatics* 24:1401–1402
62. Ishida T, Kinoshita K (2007) PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res* 35(Web Server issue):W460–W464
63. Peng K et al (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* 7:208
64. Huang YJ, Acton TB, Montelione GT (2014) DisMeta: a meta server for construct design and optimization. *Methods Mol Biol* 1091:3–16
65. Kozlowski LP, Bujnicki JM (2012) MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics* 13:111
66. Mizianty MJ et al (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics* 26:i489–i496
67. Xue B et al (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim Biophys Acta* 1804:996–1010
68. Schlessinger A et al (2009) Improved disorder prediction by combination of orthogonal approaches. *PLoS One* 4:e4433
69. Monastyrskyy B et al (2011) Evaluation of disorder predictions in CASP9. *Proteins* 79 Suppl 10:107–118
70. Peng ZL, Kurgan L (2012) Comprehensive comparative assessment of in-silico predictors of disordered regions. *Curr Protein Pept Sci* 13:6–18
71. Gsponer J et al (2008) Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science* 322:1365–1368
72. Pajkos M et al (2012) Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. *Mol Biosyst* 8:296–307
73. Kovacs E et al (2010) Dual coding in alternative reading frames correlates with intrinsic protein disorder. *Proc Natl Acad Sci U S A* 107:5429–5434
74. Graslund S et al (2008) Protein production and purification. *Nat Methods* 5:135–146
75. Dosztanyi Z et al (2007) Prediction of protein disorder at the domain level. *Curr Protein Pept Sci* 8:161–171
76. Oldfield CJ et al (2013) Utilization of protein intrinsic disorder knowledge in structural proteomics. *Biochim Biophys Acta* 1834:487–498

77. Suskiewicz MJ et al (2011) Context-dependent resistance to proteolysis of intrinsically disordered proteins. *Protein Sci* 20: 1285–1297
78. Linding R et al (2004) A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. *J Mol Biol* 342: 345–353
79. Esnouf RM et al (2006) Honing the in silico toolkit for detecting protein disorder. *Acta Crystallogr D Biol Crystallogr* 62: 1260–1266
80. Hogg-Johnson S et al (2012) A randomised controlled study to evaluate the effectiveness of targeted occupational health and safety consultation or inspection in Ontario manufacturing workplaces. *Occup Environ Med* 69:890–900
81. Oldfield CJ et al (2005) Addressing the intrinsic disorder bottleneck in structural proteomics. *Proteins* 59:444–453
82. Grabowski M et al (2007) Structural genomics: keeping up with expanding knowledge of the protein universe. *Curr Opin Struct Biol* 17:347–353
83. von Heijne G (2007) The membrane protein universe: what's out there and why bother? *J Intern Med* 261:543–557
84. Kloppe E, Punta M, Rost B (2012) Structural genomics plucks high-hanging membrane proteins. *Curr Opin Struct Biol* 22:326–332
85. Love J et al (2010) The New York Consortium on Membrane Protein Structure (NYCOMPS): a high-throughput platform for structural genomics of integral membrane proteins. *J Struct Funct Genomics* 11:191–199
86. Kelly L et al (2009) A survey of integral alpha-helical membrane proteins. *J Struct Funct Genomics* 10:269–280
87. Pieper U et al (2013) Coordinating the impact of structural genomics on the human alpha-helical transmembrane proteome. *Nat Struct Mol Biol* 20:135–138
88. Bernado P et al (2005) A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc Natl Acad Sci U S A* 102:17002–17007
89. Schneider R et al (2012) Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy. *Mol Biosyst* 8:58–68
90. Fisher CK, Stultz CM (2011) Constructing ensembles for intrinsically disordered proteins. *Curr Opin Struct Biol* 21:426–431
91. Varadi M et al (2014) pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res* 42:D326–D335
92. Huang F et al (2012) Subclassifying disordered proteins by the CH-CDF plot method. *Pac Symp Biocomput*:128–139
93. Xue B et al (2009) CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. *FEBS Lett* 583:1469–1474
94. Altschul SF et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
95. Wootton JC (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* 18:269–285
96. Meszaros B, Dosztanyi Z, Simon I (2012) Disordered binding regions and linear motifs—bridging the gap between two models of molecular recognition. *PLoS One* 7:e46829
97. Meszaros B, Simon I, Dosztanyi Z (2009) Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol* 5:e1000376
98. Mooney C et al (2012) Prediction of short linear protein binding regions. *J Mol Biol* 415:193–204
99. Disfani FM et al (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics* 28:i75–i83
100. Dinkel H et al (2014) The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Res* 42:D259–D266
101. Davey NE et al (2012) SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. *Nucleic Acids Res* 40: 10628–10641
102. Nguyen Ba AN et al (2012) Proteome-wide discovery of evolutionary conserved sequences in disordered regions. *Sci Signal* 5:rs1
103. Bryson K, Cozzetto D, Jones DT (2007) Computer-assisted protein domain boundary prediction using the DomPred server. *Curr Protein Pept Sci* 8:181–188
104. Kim DE et al (2005) Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. *Proteins* 61 Suppl 7:193–200
105. Hunter S et al (2012) InterPro in 2011: new developments in the family and domain

- prediction database. *Nucleic Acids Res* 40(Database issue):D306–D312
106. Finn RD et al (2014) Pfam: the protein families database. *Nucleic Acids Res* 42(Database issue):D222–D230
107. Letunic I, Doerks T, Bork P (2012) MART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res* 40(Database issue):D302–D305
108. Sharma S et al (2009) Construct optimization for protein NMR structure analysis using amide hydrogen/deuterium exchange mass spectrometry. *Proteins* 76:882–894
109. Pantazatos D et al (2004) Rapid refinement of crystallographic protein construct definition employing enhanced hydrogen/deuterium exchange MS. *Proc Natl Acad Sci U S A* 101:751–756
110. Zhang Y, Stec B, Godzik A (2007) Between order and disorder in protein structures: analysis of “dual personality” fragments in proteins. *Structure* 15:1141–1147
111. Fuxreiter M et al (2004) Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J Mol Biol* 338:1015–1026
112. Brown CJ et al (2011) Evolution and disorder. *Curr Opin Struct Biol* 21:441–446
113. Akiva E et al (2012) A dynamic view of domain-motif interactions. *PLoS Comput Biol* 8:e1002341
114. Mittag T, Kay LE, Forman-Kay JD (2010) Protein dynamics and conformational disorder in molecular recognition. *J Mol Recognit* 23:105–116
115. Mitrea DM, Kriwacki RW (2012) Cryptic disorder: an order-disorder transformation regulates the function of nucleophosmin. *Pac Symp Biocomput*:152–163
116. Ward JJ et al (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics* 20:2138–2139
117. Shimizu K, Hirose S, Noguchi T (2007) POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. *Bioinformatics* 23:2337–2338
118. Kissinger CR et al (1995) Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex. *Nature* 378:641–644

Part II

Soluble Proteins

Chapter 4

Characterization and Production of Protein Complexes by Co-expression in *Escherichia coli*

Matthias Haffke, Martin Marek, Martin Pelosse, Marie-Laure Diebold, Uwe Schlattner, Imre Berger, and Christophe Romier

Abstract

The functional units within cells are often macromolecular complexes rather than single species. Production of these complexes as assembled homogenous samples is a prerequisite for their biophysical and structural characterization and hence an understanding of their function in molecular terms. Co-expression in *Escherichia coli* has been used routinely to decipher the subunit composition, assembly, and production of whole protein complexes. Such complexes can then be used to reconstitute protein/nucleic acid complexes in vitro. In this chapter we present protocols for the widely utilized ACEMBL and pET-MCN/pET-MCP vector series which enable the rapid and automated co-expression of protein complexes in *Escherichia coli*.

Key words Protein complexes, Co-expression, *Escherichia coli*, ACEMBL, pET-MCN, pET-MCP, Cloning, SLIC, Plasmid concatenation, Expression tests, High throughput, Automation

1 Introduction

Increasing cellular complexity requires functional activities that can be combined and modulated depending on the cellular state. Macromolecular complexes, where various enzymatic activities can be coupled and regulated by specific subunits, have evolved to meet these needs. However, even if the subunit composition of a protein complex is known, understanding precisely how these are assembled still represents a major challenge for biophysicists and structural biologists. Among the technologies developed to address this issue, co-expression of subunits as recombinant proteins has emerged as the method of choice for reconstituting protein complexes, combining the strengths of in vitro (choice of the constructs to be used for each subunit) and in vivo (complex reconstitution in vivo) approaches. As for single protein expression, the ease of the genetic manipulation,

the rapid tunable growth, and the low cost handling of *Escherichia coli* have made these bacteria the most common and favored co-expression hosts [1, 2].

Co-expression can be performed in *E. coli* using different strategies based on the use of single vectors or several vectors that are co-transformed. Each vector may harbor one or several promoters that control one or several genes in single-gene expression cassettes or polycistrons [3–5]. Numerous co-expression systems based on these strategies, and each with their own merits, have been developed to enable co-expression of proteins in *E. coli* [5–17]. Importantly, proof-of-concept experiments using different test cases have shown that outcomes are complex dependent [8, 9, 18]. Specifically, these studies have demonstrated that in order to obtain homogeneous and soluble protein complexes for biochemical and structural studies, it requires the evaluation of a large number of constructs in which the protein boundaries are varied. In addition, assessment of the position of the purification tag (protein bearing the tag, N- or C-terminal tag location) is also crucial for successful complex reconstitution to avoid false-negatives or to address problems of stoichiometry [8].

Therefore, the production of protein complexes, particularly those with many subunits, requires a large combinatorial set of experiments to be carried out which in turn necessitates high-throughput (HT) approaches and semi-automation or full automation of co-expression analyses. In this context, two co-expression systems for multi-subunit complexes have been developed and successfully automated, namely, ACEMBL and pET-MCN/pET-MCP technologies [7, 8, 19].

Both ACEMBL and pET-MCN/pET-MCP systems follow a similar workflow for co-expression analyses (Fig. 1). Specifically, both systems make use of initial vectors that harbor a single gene or expression element (such as a polycistron). These initial vectors are then joined together into one plasmid harboring all genes encoding the subunits of the complex under investigation. Although vector assembly can be carried out with both systems either prior or subsequent to performing the initial test co-expression experiments, the ACEMBL strategy is geared to the use of a merged multi-expression vector for the initial co-expression experiments. By contrast, the pET-MCN/pET-MCP strategy favors co-expression by using co-transformed single vectors and the subsequent merging of the plasmids once initial experiments have been successfully interpreted [7–9].

A further difference between the ACEMBL and pET-MCN/pET-MCP systems exists in the way of merging vectors. Specifically, for ACEMBL vector assembly (concatemerization; Fig. 2), several plasmids, each with a different resistance marker, are fused together by Cre-LoxP recombination using the Cre recombinase. Since the Cre-LoxP fusion is an equilibrium reaction, all possible multi-gene fusions are made simultaneously by Cre recombinase, and the

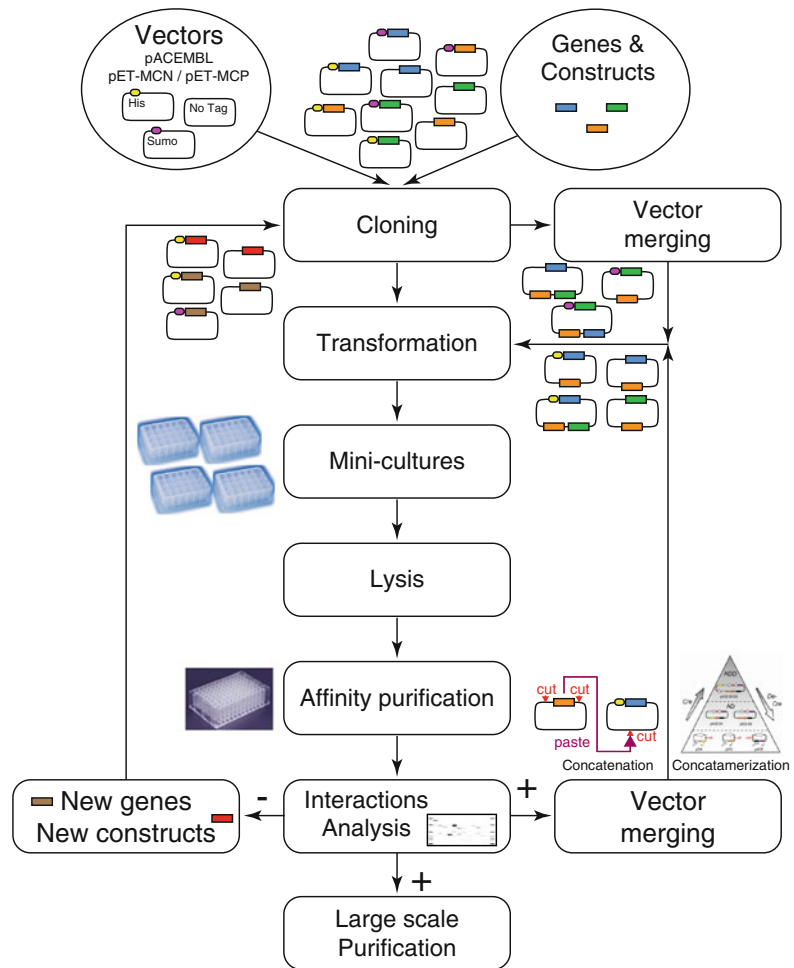


Fig. 1 Flowchart of co-expression experiments for protein complex assembly, deciphering, and reconstitution by the ACEMBL and pET-MCN/pET-MCP systems. Genes and constructs are depicted as *colored rectangles*, whereas vectors are shown as *rectangles with rounded corners*. Affinity/fusion tags are shown on the vectors as *colored circles*. The flowchart applies to both pET-MCN/pET-MCP and ACEMBL co-expression systems, though the details for some of the steps are different. Please refer to the text for the description of these differences

fusions containing the desired gene combinations can be selected for by challenging transformants with the appropriate combination of antibiotics. On the other hand, pET-MCN/pET-MCP vector merging (concatenation; Fig. 3) currently relies on conventional restriction/ligation procedures (*see Note 1*) [8].

In the next sections, the materials and methods used to perform co-expression experiments in *E. coli* using the flowchart shown in Fig. 1 are described. Unless stated explicitly, the material and methods used apply equally to the ACEMBL and pET-MCN/pET-MCP co-expression systems.

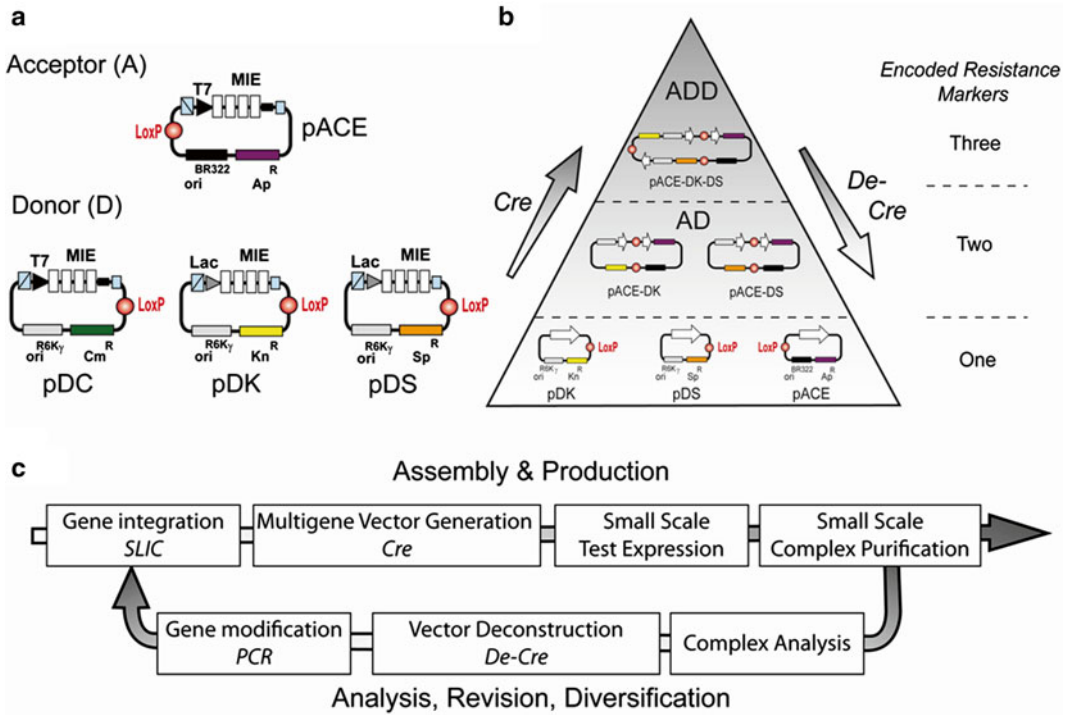


Fig. 2 ACEMBL system for multiprotein production in *E. coli*. **(a)** Donor and acceptor vectors contain *LoxP* sequences and a multiple insertion elements (MIE) for inserting one or several genes of interest. Acceptors have regular replicons (BR322). Donors have a conditional origin of replication derived from R6K γ phage. Promoters (T7, *lac*), terminators (black squares), and homing endonuclease sites (I-CeuI and PI-SceI, light-blue boxes, strike-through) and matching *Bst*XI sites (small light-blue squares) are shown (for further details, see [7]). Antibiotic resistance makers are as follows: *Ap* ampicillin, *Cm* chloramphenicol, *Kn* kanamycin, and *Sp* spectinomycin. Genes of interest are inserted into acceptor or donor vectors into the MIEs by SLIC. **(b)** Incubation of acceptor and donor constructs (genes shown as white arrows) with Cre recombinase results in all combinations of fusions including acceptor-donor (AD) and acceptor-donor-donor (ADD) fusions. Each fusion is quasi barcoded by the resistance marker combination. **(c)** The ACEMBL HT pipeline. Genes are integrated by ligation-independent methods (SLIC) followed by combinatorial multi-gene vector generation using Cre-*LoxP* fusion (tandem recombineering, TR), followed by protein expression and analysis of purified complex. Genes in donors and acceptors can be modified iteratively and the multi-gene expression constructs reassembled by TR [7, 19]

2 Materials

If not stated otherwise, all solutions are prepared with ultrapure water (Millipore Milli-Q system) and analytical grade reagents. Buffers, enzymes, and 1,000 \times antibiotic stock solutions are stored at -20°C .

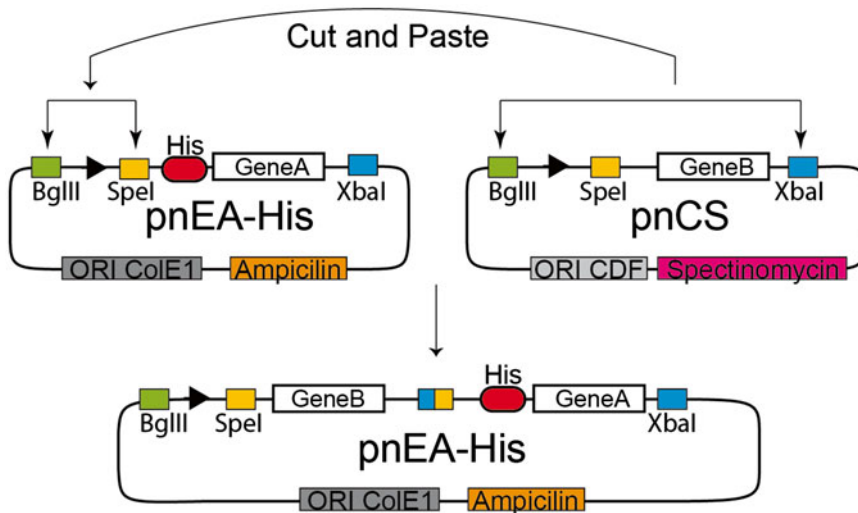


Fig. 3 Concatenation procedure for the pET-MCN/pET-MCP vectors. Example of the concatenation of two vectors of the pET-MCN vector series: pEA-His (vector encoding an N-terminal poly-histidine tag in front of protein A) and pCS (vector expressing the native protein B). The pEA-His (acceptor vector) is linearized by removing with the restriction enzymes BglII and SpeI part of its promoter (T7 promoter and lacO). The full promoter of the pCS (donor vector) is cut out with the restriction enzymes BglII and XbaI. After ligation of the pCS promoter with the linearized pEA-His vector, a new vector is obtained based on the backbone of the pEA-His vector and whose promoter controls the genes of proteins A and B. More information on the pET-MCN/pET-MCP vector series can be obtained at http://archive.igbmc.fr/recherche/Sup_papers/Cavarelli/pET-MCN/index.html and [8]

2.1 Sequence- and Ligation-Independent Cloning: SLIC (pET-MCN/pET-MCP)

2.1.1 Linearization of the Vector by Restriction Digestion

1. NdeI and BamHI restriction enzymes (e.g., New England Biolabs).
2. 10× NEBuffer 3 (New England Biolabs).
3. Shrimp alkaline phosphatase (Takara).
4. NucleoSpin Gel and PCR Cleanup Kit (Macherey-Nagel).
5. NanoDrop 1000 (Thermo Scientific).

2.1.2 Insert Preparation by PCR

1. Phusion High-Fidelity DNA Polymerase (New England Biolabs).
2. 5× Phusion HF Buffer (included in kit).
3. 10 mM dNTP mix (New England Biolabs).
4. 100 μM SLIC primers (*see Note 2*).
5. Thermocycler (e.g., Thermocycler T3000, Biometra).

2.1.3 Insert Purification

1. Agarose gel purification system (e.g., Maxigel ECO2, Apelex).
2. 50× TAE buffer: 0.04 M Tris-acetate, 1 mM EDTA (pH 8.0). Weigh 242 g Tris base (MW: 121.10 g/mol) and dissolve it in 600 ml H₂O in a 1-l graduated cylinder. Add 57.1 ml of acetic

acid (100 %) and 100 ml of 0.5 M EDTA, pH 8.0. Fill up to a total volume of 1 l with H₂O. Filter through 0.22 µm filter. Store at room temperature.

3. Agarose type D-5 DNA grade.
4. 6× DNA loading dye.
5. GeneRuler DNA Ladder Mix (Thermo Scientific).
6. Safe Imager 2.0 Blue Light Transilluminator (Life Technologies).

2.1.4 T4 DNA

Polymerase Treatment

1. T4 DNA polymerase (e.g., New England Biolabs).
2. 10× NEBuffer 2 (included in kit).
3. 10× BSA (100× BSA included in kit; dilute 1:10 with H₂O).

2.2 SLIC and Self-SLIC (ACEMBL)

2.2.1 Linearization of the Vector and Insert Preparation by PCR

1. Phusion High-Fidelity DNA Polymerase (New England Biolabs).
2. 5× Phusion HF Buffer or 5× Phusion GC Buffer (included in kit).
3. DMSO (included in kit).
4. 10 mM dNTP mix (New England Biolabs).
5. 100 µM SLIC primers (*see Note 2*).
6. Thermocycler (e.g., Thermocycler T3000, Biometra).

2.2.2 DpnI Digest

1. DpnI (e.g., New England Biolabs).
2. 10× NEBuffer 4 (included in kit).
3. QIAquick Gel Extraction Kit (Qiagen).
4. Agarose gel electrophoresis system (e.g., Mini-Sub Cell GT System, Bio-Rad).
5. 5× TBE Buffer: 0.89 M Tris base, 0.89 M boric acid, 20 mM EDTA (pH 8.0). Weigh 108 g Tris base (MW: 121.10 g/mol) and 55 g boric acid (MW: 61.83 g/mol) and add 40 ml of 0.5 M EDTA, pH 8.0 in a 2-l graduated cylinder. Fill up to a total volume of 2 l with H₂O. Filter through 0.22 µm filter and autoclave to prevent precipitation during long-term storage. Store at room temperature.
6. Agarose type D-5 DNA grade.
7. 6× DNA loading dye: 30 % (v/v) glycerol, 0.125 % (w/v) bromophenol blue, 0.125 % (w/v) xylene cyanol FF, 0.125 % (w/v) orange G (*see Note 3*).
8. 1 kb DNA ladder and 100 bp DNA ladder (New England Biolabs).

2.2.3 T4 DNA

Polymerase Treatment

1. T4 DNA polymerase (New England Biolabs).
2. 10× NEBuffer 2 (included in kit).
3. 10× BSA (100× BSA included in kit; dilute 1:10 with H₂O).
4. 10 mM dCTP solution (*see Note 4*).

2.2.4 SLIC/Self-SLIC Annealing Step

1. 37 °C heat block for SLIC/98 °C heat block for self-SLIC.
2. 10× T4 DNA ligase buffer (e.g., New England Biolabs).

2.3 Transformation of Chemical- Competent Cells

1. DH5α (pET-MCN/pET-MCP) and BW23474 (ACEMBL; *see Note 5*) chemical-competent cells or equivalent for cloning.
2. BL21[DE3] chemical-competent cells for expression.
3. LB broth (Miller).
4. LB agar (Miller).
5. Required antibiotics (1,000× stock solutions): ampicillin 50 mg/ml (in H₂O), chloramphenicol 34 mg/ml (in 100 % ethanol), spectinomycin 50 mg/ml (in H₂O), tetracycline 12.5 mg/ml (in 70 % ethanol), gentamicin 10 mg/ml (in H₂O), kanamycin 30 mg/ml (in H₂O) (*see Note 6*).

2.4 Colony PCR

1. Taq polymerase (e.g., New England Biolabs).
2. 10× Taq buffer (included in kit).
3. 10 mM dNTP mix (New England Biolabs).
4. 100 μM gene-specific primers (e.g., SLIC primers).
5. Material for agarose gel analysis (*see Subheading 2.1*).

2.5 Cre-LoxP Recombination (ACEMBL)

1. Cre recombinase (New England Biolabs).
2. 10× Cre recombinase reaction buffer (New England Biolabs).

2.6 Mini-cultures

1. Medium M1: for each mini-culture to be done (*see Note 7*), mix 1.9 ml 2× LB (LB medium concentrated twice) with 100 μl of 10 % (w/v) glucose stock solution (final glucose concentration: 0.5 %) (*see Note 8*). Supplement this medium with the required antibiotics.
2. Medium M2: for each mini-culture to be done (*see Note 7*), mix 1.86 ml 2× LB with 100 μl 12 % (w/v) lactose stock solution (final lactose concentration: 0.6 %) (*see Note 8*), 40 μl of a 1 M HEPES pH 7.0 stock solution (final HEPES concentration: 20 mM), and 1 μl of a 1 M IPTG stock solution (EUROMEDEX; final IPTG concentration: 0.5 mM). Supplement this medium with the required antibiotics.
3. 24-well deep-well plates (Whatman).
4. Gas-permeable seals (e.g., BREATHseals Greiner).

2.7 Lysis

1. Lysis buffer(s): these buffers are project dependent and should be chosen by the investigator. Prepare 1.5 ml of resuspension buffer per mini-culture test. Some additional resuspension buffer (a few ml) should also be prepared for equilibrating the affinity resin (*see Subheading 3.10*).
2. Sonicator (e.g., 4-head Vibracell 75043 sonicator, Bioblock Scientific).

2.8 Affinity Purification

1. Affinity resin (e.g., TALON affinity resin, Clontech, for poly-histidine tags). The affinity resin chosen is dependent on the affinity tag used (*see Note 9*). The volume of affinity resin slurry used for each mini-culture test should be chosen so that the final volume of beads, once the supernatant has been removed, is approximately 15 μ l (e.g., for an affinity resin slurry composed of 50 % beads and 50 % buffer, use 30 μ l of affinity resin slurry per mini-test).
2. Wash buffers: as for the lysis buffers, these buffers are project dependent (*see Note 10*). Prepare 1 ml of washing buffer per mini-culture test.
3. 96-well deep-well plate(s) (e.g., Whatman).
4. Pipetting robot (e.g., Freedom EVO, Tecan) or multichannel pipette (e.g. Impact2, Matrix).
5. SDS-PAGE system (e.g., Mini-PROTEAN, Bio-Rad) or microfluidics capillary electrophoresis system (e.g., LabChip GXII, Caliper Life Sciences).

2.9 Vector Concatenation (pET-MCN/pET-MCP)

1. Restriction enzymes: BglII, SpeI, and XbaI (optionally NheI or AvrII replacing XbaI) (*see Note 11*).
2. Shrimp alkaline phosphatase (Takara).
3. 10 \times NEBuffer 2 (New England Biolabs).
4. NucleoSpin Gel and PCR Cleanup Kit (Macherey-Nagel).
5. T4 DNA ligase (EpiCentre).

3 Methods

3.1 SLIC into pET-MCN/pET-MCP Vectors

The pET-MCN and pET-MCP vectors offer a large variety of affinity purification tags and fusions to be used in co-expression experiments (*see Note 12*). To facilitate initial cloning, all these vectors have the same multiple cloning sites (*see Note 13*). These sites can be used for cloning by conventional restriction/ligation methods. However, sequence- and ligation-independent cloning (SLIC) is preferred since it is more versatile and efficient [20, 21] (Fig. 4). Therefore, SLIC into pET-MCN/pET-MCP vectors will only be described in detail. Both strategies (restriction/ligation and SLIC) use the same method for vector linearization (*see Note 14*).

3.1.1 Linearization of the Vector (Conventional Restriction/Ligation and SLIC)

1. In a 1.5 ml Eppendorf tube, mix 5 μ g of vector with 5 μ l of NEBuffer 3. Make up to 50 μ l with H₂O. Add 5 U of NdeI and 5 U of BamHI restriction enzymes and 2 U of shrimp alkaline phosphatase. Incubate at 37 °C for 2 h.
2. Purify the vector using the NucleoSpin Gel and PCR Cleanup Kit following the manufacturer's instructions.

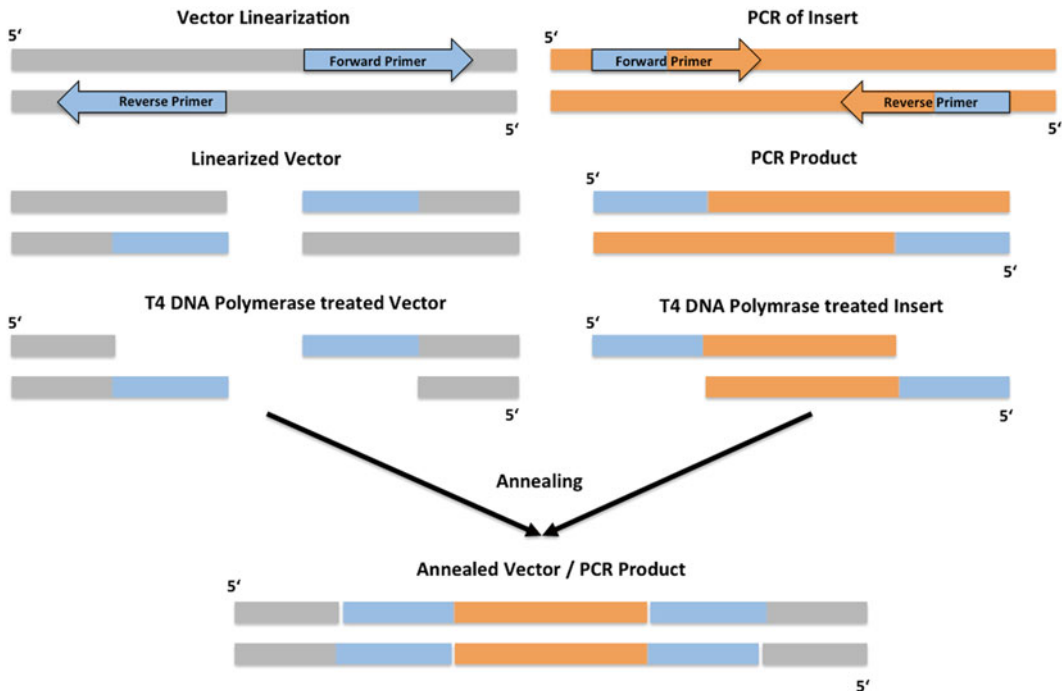


Fig. 4 Schematic representation of the SLIC procedure. The linearized vector is shown on the *left*, and the PCR product of the insert amplification is shown on the *right*. Vector linearization by PCR is shown (*see* Subheading 3.2) but can also be performed by restriction cutting (*see* Subheading 3.1). 5' overhangs are generated by T4 DNA polymerase treatment. Linearized vector and PCR product are joined in an annealing reaction via the SLIC region (shown in *blue*). The annealed product is transformed into *E. coli* cells, and the double-strand break is repaired in vivo by the DNA repair machinery

3. Estimate the vector concentration by absorbance at 260 nm (e.g., NanoDrop 2000, Thermo Scientific).
4. Store the purified vector at -20°C for subsequent SLIC reactions.

3.1.2 Preparation of the Insert by PCR

1. Set up a 50 μl PCR reaction in a 0.5 ml PCR tube: mix 1 μl template DNA (approximately 50 ng) with 10 μl 5 \times Phusion HF, 1 μl 10 mM dNTP mix, 1 μl of forward primer, and 1 μl of reverse primer, and fill up to 49.5 μl with H_2O . Add 1 U of Phusion High-Fidelity DNA Polymerase and mix.
2. Perform the PCR reaction. Typically, templates are initially denatured at 98°C for 60 s. This initial cycle is followed by 40 cycles, starting with a denaturing step at 98°C for 10 s, the specific annealing temperature for 15 s and the elongation step at 72°C for 15 s (for 1 kb product size). After cycling, a single final step at 72°C for 5 min is performed followed by cooling to 15°C and holding at this temperature.

3. Mix the PCR reaction with 10 μ l 6 \times DNA loading dye, load on a 0.5–2 % TAE agarose gel (depending on the size of the PCR product), and run the gel at 200 V until the PCR products are well resolved.
4. Excise the band corresponding to the PCR products using the Safe Imager 2.0 Blue Light Transilluminator (*see* **Note 15**) and transfer it to a 1.5 ml sterile Eppendorf tube.
5. Extract the DNA from the gel slices using the NucleoSpin Gel and PCR Cleanup Kit.
6. Determine the concentration of the extracted DNA by absorbance at 260 nm.

3.1.3 SLIC Reaction

1. In a 1.5 ml Eppendorf tube, prepare a 10 μ l SLIC reaction by mixing 50 ng of linearized vector, 100 ng of insert, 1 μ l of NEBuffer 2, and 1 μ l of 10 \times BSA and complete to 9.75 μ l with H₂O.
2. Add 1 U of T4 DNA polymerase to start the reaction.
3. Immediately start thawing DH5 α cloning-dedicated chemical-competent cells in an ice bucket (*see* Subheading 3.4 for the transformation protocol).
4. After 10 min of SLIC reaction, transfer 3 μ l of the reaction to a 1.5 ml Eppendorf tube for transformation of the DH5 α chemical-competent cells (*see* Subheading 3.4 for the transformation protocol) and freeze the rest of the reaction.

3.2 SLIC into ACEMBL Vectors

Sequence- and ligation-independent cloning (SLIC) is a fast and reliable method to insert genes of interest into the ACEMBL vectors. It is advisable to use the same SLIC annealing region for various inserts. In this way one can standardize the ACEMBL vector library and prepare stocks of linearized vector for annealing with various inserts. The SLIC process is shown in Fig. 4. If gene insertion targets the multiple insertion element (MIE), ACEMBL vectors can be linearized as described in Subheading 3.1 by simple restriction enzyme digestion using specific restriction enzymes matching those of the MIE. This alleviates the requirement to order primers and to run PCR on the empty vector.

3.2.1 Linearization of the Vector and Insert Preparation by PCR

1. Set up PCR reactions for the insert and the vector separately.
2. For the insert, set up a 50 μ l PCR reaction in a 0.5 ml PCR tube: mix 1 μ l template DNA (approximately 10 ng) with 10 μ l 5 \times Phusion HF or 10 μ l 5 \times Phusion GC Buffer (*see* **Note 16**), 1 μ l 10 mM dNTP mix, 1 μ l of forward primer, and 1 μ l of reverse primer, and make up to 49.5 μ l with water. Add 1 U Phusion High-Fidelity DNA Polymerase and mix.
3. Set up the same reaction for the vector (*see* **Note 17**).

4. Choose appropriate annealing temperatures for the specific primers chosen to perform the PCR (*see Note 18*). Typically, templates are initially denatured at 98 °C for 60 s; followed by 30 cycles, starting with a denaturing step at 98 °C for 20 s, the specific annealing temperature for 30 s and the elongation step at 72 °C for 15 s (for 1 kb product size). After cycling, a single final step at 72 °C for 5–10 min with subsequent cooling to 4 °C and pausing at that temperature is performed.

3.2.2 DpnI Digest and Purification of PCR Product and Linearized Vector

1. To digest the template present in the PCR reactions, add 20 μ l of DpnI directly to the 50 μ l PCR product and incubate at 37 °C for 2 h (*see Note 19*).
2. Mix with 10 μ l 6 \times DNA loading dye, load on 0.5–2 % TBE agarose gel (depending on size of the PCR product), and run the gel at 100–120 V until the bands of the DNA ladder corresponding to the size of the PCR products (amplified insert and linearized vector) are well resolved.
3. Excise the band corresponding to the PCR products (amplified insert and linearized vector) using a UV light box and transfer to a 2 ml sterile Eppendorf tube (*see Notes 15 and 20*).
4. Extract the DNA from the gel slices using the QIAquick Gel Extraction Kit following the instructions in the product's manual.
5. Determine the concentration of the extracted DNA by absorbance at 260 nm.

3.2.3 T4 DNA Polymerase Treatment of PCR Product and Linearized Vector

1. The PCR product and the linearized vector are treated independently with T4 DNA polymerase. Set up the reaction in a 0.5 ml PCR tube: mix 2 μ l 10 \times NEBuffer 2, 2 μ l of 10 \times BSA, 0.5 U T4 DNA polymerase, and 1 μ g of the purified PCR product or linearized vector (*see Note 21*) in a total volume of 20 μ l. For a 20 bp overhang between PCR product and vector, incubate for 30 min at room temperature (*see Note 22*).
2. Stop the reaction by adding 2 μ l 10 mM dCTP and store on ice.

3.2.4 SLIC Annealing

1. Set up the annealing reaction by mixing 200 ng of the T4 DNA polymerase-treated vector with a 1:3 molar ratio of T4 DNA polymerase-treated insert in a 10 μ l reaction containing 1 μ l 10 \times T4 DNA ligase buffer (*see Note 23*).
2. Use 5 μ l of the SLIC annealing reaction to transform cloning-dedicated BW23474 chemical-competent cells (*see Subheading 3.4 for the transformation protocol*).

3.3 Self-SLIC Method to Modify Genes of Interest

Self-SLIC can be used to introduce mutations, such as insertions, deletions, or codon changes into a gene of interest. The primer design for all three types of gene modification is shown in Fig. 5. By taking advantage of the degeneracy of the genetic code, restriction sites can be inserted or deleted in the SLIC region of the designed

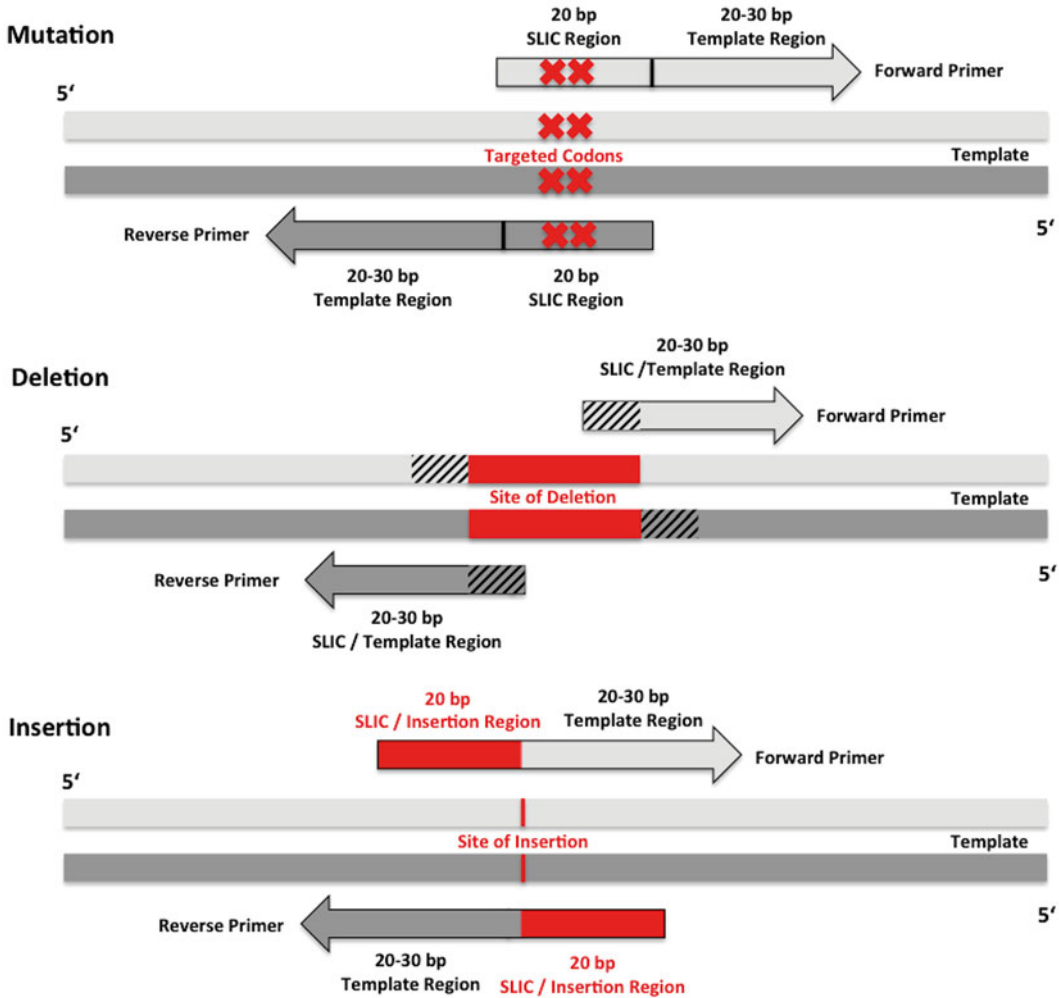


Fig. 5 Primer design for gene modification in ACEMBL vectors. *Top:* primer design for site-directed mutagenesis to exchange codons within a 20 bp sequence of choice. The targeted codons are indicated with *red crosses*. The complete 20 bp SLIC region can be used for mutations. Care should be taken that the template region is longer and has a higher melting temperature than the 20 bp SLIC region to avoid unspecific side products in the PCR reaction. *Middle:* primer design to introduce deletions. The deleted region within the template is indicated in *red*. Note that the primers carry a combined SLIC region, utilizing 10 bp at the 5' site and 10 bp at the 3' site of the deleted sequence, resulting in a 20 bp overall SLIC region. *Bottom:* primer design for inserting additional sequences up to 20 bp at a targeted site. The site of insertion is indicated with a *red line* on the template double strand. Care should be taken that the template region is longer and has a higher melting temperature than the 20 bp SLIC/insertion region to avoid primer-dimer formation or unspecific PCR products

primers. This allows, by simple restriction digest, the rapid identification of clones that contain a desired modification. After DpnI digestion for template removal and T4 DNA polymerase treatment of the linearized vector, the linearized vector is denatured at 98 °C, annealed, and transformed into *E. coli* cells for plasmid propagation. Vector linearization of the vector by PCR, DpnI digestion, and T4 DNA polymerase treatment are carried out as described in Subheading 3.2.

3.3.1 Self-SLIC Annealing

1. Set up the annealing reaction using 300 ng of the T4 DNA polymerase-treated linearized vector in a 10 µl reaction containing 1 µl 10× T4 DNA ligase buffer.
2. Denature the linearized vector by heating to 98 °C for 5 min.
3. Allow to cool down slowly to room temperature.
4. Use 5 µl of the Self-SLIC annealing reaction to transform cloning-dedicated BW23474 chemical-competent cells (*see* Subheading 3.4).

3.4 Transformation of Cloning-Dedicated Chemical-Competent Cells

1. Mix the desired amount of SLIC (pET-MCN/pET-MCP) or SLIC/Self-SLIC annealing (ACEMBL) reaction with 50 µl of DH5α (pET-MCN/pET-MCP) or BW23474 (ACEMBL) chemical-competent cells and incubate on ice for 30 min, heat shock at 42 °C for 45 s, incubate on ice for 2 min, add 400 µl of LB broth, and incubate in a 37 °C shaker for 1 h. For the ACEMBL system, use *pir*-negative strains for acceptor vectors and *pir*-positive strains for donor vectors.
2. Pellet the cells by centrifugation at 2,500 × *g* for 2 min, take off 300 µl supernatant, and resuspend the pellet in the remaining 150 µl. Plate 100 µl of this concentrated cell suspension on a LB agar plate with the appropriate antibiotics. Dilute the remaining 50 µl of cells 1:10 with LB broth and plate 100 µl of this diluted cell suspension on a second LB agar plate with the appropriate antibiotics.

3.5 Colony PCR

1. Set up a 20 µl PCR reaction in a 0.5 ml PCR tube: mix 1.5 µl of *E. coli* culture with 2 µl 10× Taq buffer, 1 µl 10 mM dNTP mix, 1 µl of gene-specific forward primer, and 1 µl of gene-specific reverse primer, and fill up to 19.5 µl with H₂O. Add 1 U of Taq and mix.
2. Perform the PCR reaction. Typically, cells are disrupted and templates are denatured at 98 °C for 180 s. This initial cycle is followed by 50 cycles, starting with a denaturing step at 98 °C for 30 s, the specific annealing temperature for 30 s, and the elongation step at 72 °C for 1 min (for 1 kb product size). After cycling, a single final step at 72 °C for 5 min with subsequent cooling to 15 °C and pausing at that temperature is performed.

3. Mix the PCR reaction with 10 μ l 6 \times DNA loading dye, load on a 0.5–2 % TAE agarose gel (depending on the size of the PCR product), and run the gel at 200 V until the PCR products are well resolved.
4. Analyze the gel under UV light, looking for PCR fragments corresponding to the size of the inserts.
5. For the cultures showing an insert at the right size, purify the plasmid using the NucleoSpin Plasmid Kit according to the manufacturer's instructions.
6. Sequence the plasmid to check that no mutations are present in the insert.

**3.6 Cre-LoxP
Recombination
of Acceptors
and Donors (ACEMBL
Vectors)**

1. Combine 1 μ g of acceptor (pACE) in a 1:1 molar ratio with one, two, or three donors of choice (pDC, pDK, pDS) in a 10 μ l reaction containing 1 μ l 10 \times Cre recombinase reaction buffer and 1 U Cre recombinase.
2. Incubate at 37 °C for 1 h (*see Note 24*).
3. Mix 5 μ l of the Cre-LoxP reaction with 50 μ l of BW23474 chemical-competent cells on ice and incubate for 30 min, heat shock at 42 °C for 45 s, incubate on ice for 2 min, add 400 μ l of LB broth, and incubate in a 37 °C shaker overnight (*see Note 25*).
4. Pellet the cells by centrifugation at 2,500 $\times g$ for 2 min, remove 300 μ l supernatant, and resuspend the pellet in the remaining 150 μ l. Plate 100 μ l of this concentrated cell suspension on a LB agar plate with the appropriate antibiotics. Dilute the remaining 50 μ l of cells 1:10 with LB broth and plate 100 μ l of this diluted cell suspension on a second LB agar plate with the appropriate antibiotics.
5. Inoculate 5 ml of LB broth containing appropriate antibiotics in a 15 ml Falcon tube from a single colony of *E. coli* cells grown on selective LB agar plates for the desired acceptor-donor fusion plasmids. Incubate at 37 °C, agitating at 150 rpm for 12 h.
6. Centrifuge the Falcon tubes for 10 min at 5,000 $\times g$ at 4 °C. Take off the supernatant and invert the Falcon tubes to drain.
7. Perform a plasmid preparation using the QIAprep Spin Miniprep Kit following the instructions in the product's manual.
8. Determine the concentration of the extracted DNA (e.g., NanoDrop 2000, Thermo Scientific).
9. Check the Cre-LoxP fusion plasmids for the presence of one copy of the acceptor and one copy of each donor used in the Cre-LoxP recombination reaction by restriction mapping (*see Note 26*).

3.7 Transformation of BL21[DE3] Expression Cells

Depending upon the strategy chosen for investigating complex formation, vectors expressing individual subunits either from one or several plasmids are co-transformed into a BL21 expression strain for co-expression:

1. Mix 50 ng of each vector with 50 μ l (*see Note 27*) of BL21[DE3] chemical-competent cells (*see Note 28*) and incubate for 30 min on ice, heat shock at 42 °C for 45 s, incubate on ice for 2 min, add 400 μ l of LB broth, and incubate in a 37 °C shaker for 1 h.
2. Pellet the cells by centrifugation at 2,500 $\times g$ for 2 min, take off 300 μ l supernatant, and resuspend the pellet in the remaining 150 μ l. Plate 100 μ l of this concentrated cell suspension on a LB agar plate with the appropriate antibiotics. Dilute the remaining 50 μ l of cells 1:10 with LB broth and plate 100 μ l of this diluted cell suspension on a second LB agar plate with the appropriate antibiotics.

3.8 Mini-cultures

Colonies growing on LB agar are used to inoculate directly co-expression mini-cultures in 24-well deep-well plates. For the mini-cultures, a specific medium is generally used initially that combines auto-induction [22] and IPTG induction strategies (*see Note 29*):

1. Add 2 ml of medium M1 per well for each mini-test (*see Note 30*) in 24-well deep-well plate(s) (*see Note 31*).
2. For inoculation, scratch from the LB agar plate with a slightly kinked pipette tip five to ten colonies (*see Note 32*) of the co-transformed *E. coli* cells, and drop the tip into the corresponding well of the 24-well deep-well plate.
3. After all wells have been inoculated, cover the 24-well deep-well plates with gas-permeable seals and transfer them to a shaker at 37 °C.
4. Allow cultures to grow for several hours at 37 °C until all cultures are cloudy (*see Note 33*).
5. Take the 24-well deep-well plates out of the shaker and let them cool down to room temperature on the bench for approximately 20 min. At the same time, lower the temperature of the shaker to 25 °C (*see Note 34*).
6. Add 2 ml of medium M2 to each well, cover the plates with gas-permeable seals, and transfer the plates to the shaker.
7. Grow the cultures at 25 °C overnight.

3.9 Lysis

1. Centrifuge the 24-well deep-well plates to pellet the cells at 1,500 $\times g$ for 10 min (*see Note 35*). Discard the spent culture media.
2. Resuspend the cell pellets in lysis buffer(s) (*see Notes 36 and 37*).

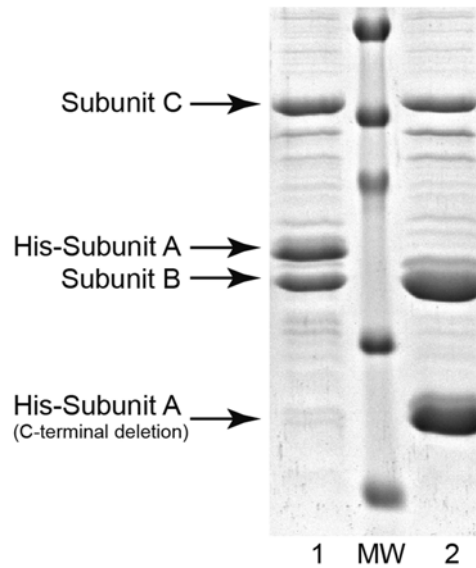


Fig. 6 SDS-PAGE analysis of co-expression experiments. Example of mini-purification of a complex of three proteins (A, B, and C) co-expressed in *E. coli*. Poly-histidine-tagged protein A is retained on the affinity resin through its tag. Proteins B and C are retained on the affinity resin through the complex they form with protein A. *Lanes 1* and *2* correspond to two different experiments where protein A was either expressed full length (*lane 1*) or C-terminally truncated (*lane 2*)

3. Lyse the cells by sonication in the 24-well deep-well plates (*see Note 38*).
4. Take 10 μ l of each lysed sample and mix it with 50 μ l of Laemmli buffer. Freeze these samples or keep them at 4 $^{\circ}$ C for analysis of overall expression after affinity purification (*see Subheading 3.10*).

3.10 Affinity Purification

The lysed samples are used for affinity purification, typically on a robotic platform [19]. However, the same protocol can be carried out with a multichannel pipette or a normal pipette (*see Note 39*). The formation of complexes from the co-expressed protein subunits is assessed by SDS-PAGE following affinity purification (Fig. 6).

1. Centrifuge the 24-well deep-well plates at $3,700 \times g$ for 20 min to remove cellular debris (*see Note 40*).
2. During centrifugation, pipette the required amount of affinity resin slurry into 1.5 ml tubes (one tube per lysis buffer and/or affinity resin type). Centrifuge the affinity resin slurry for 1 min at $200 \times g$ on a bench centrifuge. Remove the supernatant and resuspend the affinity resin in lysis buffer. The volume of lysis buffer should be twice the initial volume of affinity resin slurry.

Centrifuge the affinity resin at $200\times g$. Remove the supernatant and resuspend the affinity resin in lysis buffer and add to each well of a 96-well deep-well plate (*see* **Note 41**).

3. Transfer with a pipetting robot (or a pipette) the supernatants (e.g., 1.5 ml: *see* **Note 36**) of the centrifuged lysed samples onto the affinity resin in the 96-well deep-well plates. Cover the plates with CapMats and mix on a rocker at 4 °C for 1 h.
4. Centrifuge the 96-well deep-well plates at $200\times g$ for 10 min and remove the CapMats.
5. With a pipetting robot (or a pipette), remove the supernatants from all wells, taking care not to aspirate the affinity resin. Add 500 μ l of washing solution to each well.
6. Centrifuge the 96-well deep-well plates at $100\times g$ for 5 min.
7. Remove the supernatants and add 500 μ l of washing solution to each well as described in **step 5**.
8. Centrifuge the 96-well deep-well plates at $100\times g$ for 5 min. Remove the supernatants without adding washing buffer.
9. Add 25 μ l of Laemmli buffer (*see* **Note 42**) to each well and shake the 96 deep-well plates briefly to resuspend the affinity resin in the Laemmli buffer.
10. Incubate for 10 min the beads in the Laemmli buffer.
11. Transfer the supernatants to 96-well PCR reaction plates. The 96-well deep-well plates that were used for the binding/washing steps can be washed thoroughly with water to be reused for further co-expression mini-test purifications.
12. Load the lysis extracts (typically 5 μ l; *see* Subheading 3.9) and affinity beads samples (eluate and/or Laemmli; typically 10 μ l) on SDS-PAGE or a microfluidics capillary electrophoresis system for expression and interaction analysis. The presence of two or more co-expressed proteins in the same lane of the SDS-PAGE gel indicates that they may be associated in a complex (*see* **Note 43**).

Depending on the results of the interactions analyses, several strategies can be pursued next. In the case where no interaction is observed (*see* **Note 44**), other constructs of the proteins studied or other subunits of the complex under investigation can be considered (*see* **Note 45**).

If interaction is observed, it is possible to either continue directly to scale up (*see* **Note 46**) or to concatenate the vectors harboring the genes of the interacting proteins (*see* Subheading 3.11 for pET-MCN/pET-MCP vectors and Subheading 3.6 for ACEMBL vectors) and introduce these new vectors into the initial pool of vectors for a new cycle of mini-tests searching for new interactions and larger complexes.

3.11 Vector Concatenation (pET-MCN/pET-MCP Vectors)

1. Prepare a 20 μ l reaction composed of 2 μ l 10 \times NEBuffer 2, 2 μ g of acceptor vector (*see* **Note 47**), 2 U each of BglII and SpeI restriction enzymes, and 1 U of shrimp alkaline phosphatase (*see* **Note 48**).
2. Prepare a 20 μ l reaction composed of 2 μ l 10 \times NEBuffer 2, 2 μ g of donor vector (*see* **Note 47**), 2 U each of BglII and XbaI (alternatively NheI or AvrII; *see* **Note 11**) restriction enzymes.
3. Incubate both reactions at 37 °C for 1 h.
4. Purify on agarose gel with the NucleoSpin Gel and PCR Cleanup Kit the cut acceptor vector (*see* **Note 49**) and the gene/promoter fragment from the donor vector (*see* **Note 50**). Quantify the amount of each sample with the NanoDrop.
5. Set up a 10 μ l ligation reaction with 50 ng of acceptor vector, 100 ng of gene/promoter fragment, 1 μ l of 10 \times T4 DNA ligase buffer, and 1 U of T4 DNA ligase.
6. After 2 h (alternatively overnight) of ligation at room temperature, transform 50 μ l of DH5 α cloning-dedicated chemical-competent *E. coli* cells with 2 μ l of the ligation reaction. Plate on LB agar supplemented with the required antibiotic.
7. Set up mini-cultures (typically 4–10 ml) in 2 \times LB broth from colonies growing on LB agar in the presence of the antibiotic selecting for the acceptor vector.
8. Once the mini-cultures are cloudy, perform colony PCR reactions (*see* Subheading 3.5) to check for the presence of the gene newly inserted into the acceptor vector.
9. Purify with the NucleoSpin Plasmid Kit the plasmids from the mini-cultures that showed the presence of the inserted gene. Sequence the plasmids (*see* **Note 51**) prior to their use.

4 Notes

1. The main difference between the pET-MCN and pET-MCP vector series lies in the nature of their concatenation procedure when creating vectors harboring several genes. In the case of the pET-MCN vectors, concatenation leads to genes arranged in a polycistron that are under the control of a single promoter, whereas concatenation with the pET-MCP vectors leads to genes that are each under the control of their own promoter [8]. Both series are compatible, enabling the construction of single vectors harboring several promoters, some of them controlling the expression of one gene, whereas others control the expression of several genes. A detailed description of the

pET-MCN and pET-MCP vectors and their use can be found in [8] and on the website http://archive.igbmc.fr/recherche/Sup_papers/Cavarelli/pET-MCN/index.html.

2. The sequence- and ligation-independent cloning (SLIC) strategy makes use of inserts that include 15–20 bp sequences at their 5' and 3' ends that are homologous to the sequences found on either side of the opening made by PCR or restriction cutting of the destination vector [20]. Therefore, insert preparation requires the design of SLIC-specific primers that will code for these homologous ends and that will also prime on the gene of interest. It is also possible to introduce specific restriction sites into the primers between the homologous ends and the priming regions. This will enable standard restriction cloning into the destination vector if SLIC cloning fails or if subcloning is required in subsequent cloning steps.
3. Xylene cyanol migrates at about 4,000 bp, bromophenol blue at about 500 bp, and orange G at about 50 bp in 1 % TBE agarose gels. Use these approximate size indicators when running an agarose gel to decide the separation time.
4. Alternatively, dGTP, dATP, or dTTP can be used to inhibit the exonuclease activity of T4 DNA polymerase.
5. Use *pir*-negative strains for cloning into acceptor vectors and *pir*-positive strains for cloning into donor vectors.
6. Carbenicillin can be used in place of ampicillin (at the same concentration) to minimize the number of satellite colonies which result from the loss of antibiotic selection since carbenicillin is less susceptible to breakdown than ampicillin.
7. 4 ml of growth/expression medium is used for each mini-culture in mini-expression tests: 2 ml of medium M1 and 2 ml of medium M2.
8. The preparation of the 10 % (w/v) glucose and 12 % (w/v) lactose stock solutions may require some heating of the solutions to accelerate the dissolution process. The stock solutions should however be kept at 4 °C afterwards.
9. The use of different affinity resins for the same kind of tag (e.g., cobalt resin vs. nickel resin for poly-histidine tag) can yield different results in terms of quantity of complex bound or stoichiometry. Even the same resin from different providers can influence yields and stoichiometry. Therefore, resin type and provider should be considered as a parameter for the co-expression experiments.
10. The washing buffers for the affinity purification can be the same as the ones used for cell lysis but may also be different depending on the project.

11. The most convenient way of performing vector concatenation is to use the enzymes BglII, SpeI, and XbaI (strategy presented in the methods section). However, XbaI can be replaced with either NheI or AvrII if there are XbaI restriction sites in the genes used (*see* detailed information in [8] and the website http://archive.igbmc.fr/recherche/Sup_papers/Cavarelli/pET-MCN/index.html). Preliminary experiments show that vector concatenation for the pET-MCN/pET-MCP vectors can be carried out using SLIC.
12. The choice of the vectors is project dependent, and because of the strong influence of the affinity tag on complex formation, typically only one protein is tagged with a small affinity tag (e.g., poly-histidine tag) and all other proteins are expressed in their native form. However, this is not a rule, and a single-tag strategy may not be possible for complexes composed of many subunits. In this case, a purification strategy making use of multiple affinity tags borne by different subunit proteins is carried out to separate, at an early stage, sub-complexes from the full complex.
13. The multi-cloning sites of the pET-MCN and pET-MCP vectors are composed of the restriction sites (5'–3') NdeI, XhoI, AflII, MunI, and BamHI. Preferably, the NdeI site is used as the 5' cloning site to avoid additional N-terminal residues in the protein expressed, notably when this protein is produced in its native, untagged form.
14. Insert preparation for restriction/ligation is similar to the vector linearization described in Subheading 3.1. Inserts prepared for SLIC can be used for the restriction approach if restriction sites have been introduced into the SLIC primers (*see* **Note 2**). After amplification of the inserts by PCR, purification on agarose gel, and the estimation of their concentration, the PCR products can be digested with restriction enzymes (e.g., NdeI and BamHI) for 1 h at 37 °C and purified directly with the NucleoSpin Gel and PCR Cleanup Kit (*see* Subheading 3.1). Once purified, the inserts can be mixed in a 10 µl reaction with the linearized vector (3:1 ratio), 1 µl 10× ligation buffer, and 1 U T4 DNA ligase (EpiCentre). After 2–3 h (or overnight) ligation at room temperature, 2 µl of the ligation reaction is used for transformation of DH5α cloning-dedicated chemical-competent cells (*see* Subheading 3.4).
15. The Safe Imager 2.0 Blue Light Transilluminator does not use UV light to image the ethidium bromide-stained DNA in agarose gels avoiding UV damage, which would reduce cloning efficiency.
16. When using the Phusion High-Fidelity DNA Polymerase kit, the 5× GC buffer can help to increase the performance of

Phusion High-Fidelity DNA Polymerase on long or GC-rich templates. When working with GC-rich templates, add 3 % (v/v) DMSO as a PCR additive to aid denaturing of templates with high GC content. It is practical to run two PCR reactions with HF and GC buffer in parallel and compare yield and PCR product specificity for both reactions.

17. Use touchdown PCR protocols to avoid unspecific PCR products, especially for self-SLIC. Set the annealing temperature in the first cycle (denaturing, annealing, and extension step) 10–15 °C higher than the calculated primer-specific annealing temperature, and reduce by 1–2 °C in each subsequent complete PCR cycle until the calculated primer-specific annealing temperature is reached. Then carry out 25–30 PCR cycles (denaturing, annealing, and extension step) using this primer-specific annealing temperature. Finalize the reaction by a 10 min extension step at 72 °C.
18. When using the New England Biolabs Phusion High-Fidelity DNA Polymerase kit, calculate the annealing temperature with the manufacturer's T_m calculator tool on the website: <https://www.neb.com/tools-and-resources/interactive-tools/tm-calculator>.
19. Template digestion by DpnI is critical for successful SLIC/self-SLIC cloning to remove parent vector template which will give a high background of colonies if the target vector carries the same antibiotic resistance as the parent template. The incubation time can be extended overnight to increase efficacy, and the DpnI digest can also be performed on PCR product purified by using a PCR purification kit (i.e., QIAquick Gel Extraction Kit, Qiagen).
20. Use longer wavelengths on the UV light box (e.g., 365 nm or equivalent) and reduced light intensity to avoid any modifications to your vector/PCR product.
21. dNTPs inhibit the 3' exonuclease activity of the T4 DNA polymerase. It is therefore necessary to purify the PCR products, to remove any dNTPs from the PCR reactions beforehand.
22. Nonoverlapping overhangs between PCR product and vector can impede correct annealing, if the T4 DNA polymerase treatment time is too short.
23. Make sure that the buffer is completely dissolved. No white precipitate should be visible.
24. Longer incubation times can lead to undesired higher-molecular-weight recombination products (e.g., recombined plasmids containing multiple copies of acceptors and/or donors).
25. Long recovery times are important to obtain positive transformants when creating multiple acceptor-donor fusions due to the high selective pressure from the combination of antibiotics used.

26. Use the Cre-ACEMBLER software to create acceptor-donor fusion plasmids in silico. Using restriction enzymes cutting only donor plasmids or only acceptor plasmids can help to resolve the complex restriction pattern analysis and to identify multiple integrations of acceptors or donors. The Cre-ACEMBLER software can be found on the following webpage: http://www.embl.fr/multibac/multiexpression_technologies/cre-acemblem/index.html.
27. To enable parallel evaluation of many co-expression tests, *E. coli* transformations are performed using chemical-competent cells. Although many different protocols exist to produce chemical-competent *E. coli* cells, some of them appear incompatible with co-transformations (no transformants observed). Therefore, textbook protocols to prepare calcium-competent cells are used that are fully compatible with co-transformations. Importantly, reducing the quantity of cells used for co-transformation (typically 25 μ l rather than 50 μ l) increases the numbers of colonies that will grow on LB agar.
28. BL21[DE3] are generally used for expression with the T7 expression system. Many other *E. coli* strains that use the T7 system do however exist that can be more suitable depending on the kind of analysis to be done. These include, for instance, the co-expression of helper plasmids coding for tRNAs produced in low amounts in *E. coli* due to codon bias or strains that favor the expression of disulfide bridges containing proteins or membrane proteins. The choice of these strains is clearly project dependent and should be selected accordingly by the investigator.
29. Auto-induction medium alone appears to be suboptimal for co-expression experiments. A high cell density medium similar to the auto-induction medium (presence of glucose and lactose) but supplemented with IPTG works well in most cases. Other media can then be used such as terrific broth (TB), with standard IPTG induction. The volumes of media used depend on whether cultures are grown in 24 deep-well blocks (4 ml/well) or 50 ml Falcon tubes (15 ml/tube).
30. In case a single lysis buffer is used during purification, a single co-expression mini-test will be performed (i.e., use of a single well in a 24-well deep-well plate). If more than one lysis buffer is used, the mini-tests should be duplicated for every lysis buffer. Duplication should occur in different 24-well deep-well plates for easier handling in subsequent automated purification steps.
31. Depending on the shaker platform used, co-expression tests in 96-well deep-well plates can give poorer results than when using 24-well deep-well plates, because of poorer oxygenation of the cultures. Check whether your shaker provides optimal

- (small orbital, high-speed) shaking conditions prior to using 96-well deep-well plates for mini-cultures.
32. The yields of co-expression are higher when starting from fresh transformed cells plated on LB agar than from liquid precultures. The best is to use cells that have been transformed the day before and left at 37 °C overnight. Slightly kink a 10 µl pipette tip on the petri dish cover and then scratch carefully some colonies for inoculation.
 33. The addition of the M2 medium dilutes by a factor 2 the density of the cell cultures. Therefore, it is generally not required to make an OD₆₀₀ measurement of the cell densities prior to induction, and addition of M2 medium is carried out once all cultures are cloudy (unless some of them do not grow at all).
 34. For initial screening, use 25 °C as a “mean” temperature for co-expression, unless the project requires specific low or high temperatures for expression. This parameter can be varied in subsequent co-expression tests.
 35. Centrifugation of the 24-well deep-well plates is carried out at medium speed (1,500×g) rather than high speed to facilitate the resuspension of the cells by gentle stirring in the presence of the lysis buffer.
 36. Cells are generally resuspended in the 24-well deep-well plates in the presence of 1.5 ml of lysis buffer, this quantity being sufficient to plunge deep enough the sonication head(s) within the liquid.
 37. The choice of the lysis buffer(s) is complex dependent. In an initial test, several buffers can be used varying buffer composition and pH, additives, and salt composition and concentration. Typically, for initial testing, several salt concentrations are tested in parallel (e.g., NaCl at 50, 200, and 400 mM) at a single pH (e.g., 10 mM Tris-HCl pH 8.0). These conditions can be varied afterwards to look at the effects of other buffers, salts, and additives. In general, the large number of initial co-expression tests precludes the use of large lysis buffer screens. These can be used however afterwards in additional co-expression tests to improve the solubility of complexes that have been shown to form, albeit in suboptimal quantity and/or stoichiometry.
 38. Sonication is preferred to other lysis techniques since the yields of cell lysis are very good and this technique enables proper breaking of the DNA in small pieces that prevents handling of viscous samples in subsequent steps. Sonication can be performed with single-head or multi-head sonicators.
 39. The same protocol can also be carried out with some adaptation using 96-well filter plates. Such a protocol generally requires a specific filtration unit that may not be available in every laboratory. Filtration steps can however be carried out by centrifugation,

placing the filtration plate on the top of a 96-well deep-well plate. The speed of the centrifuge should be adapted not to break the filters.

40. Centrifugation of the 24-well deep-well plates containing the lysed samples is carried out at high speed (typically $3,700 \times g$) to avoid contamination of the supernatants for the subsequent affinity purification steps.
41. Resuspension of the affinity resin after washing it with the lysis buffer(s) should be done with a volume sufficient that it can be easily pipetted for dispatching it in the 96-well deep-well plate(s). Typically, this volume should be equal to 50 μ l of resuspension buffer times the number of mini-purifications to be done (i.e., 50 μ l of the resuspension will be dispatched in every well to be used). For dispatch, cut a 200 μ l pipette tip in its middle with scissors: this will prevent blocking the tip upon aspiration of the resuspended resin. Also pipette the resuspended resin up and down a couple of times before pipetting it for the transfer: this will keep the resuspension solution homogeneous.
42. Upon elution from the affinity column, some proteins, sub-complexes but also full complexes, can precipitate onto the affinity resin and are not found in the eluate. Analysis of the proteins bound to the affinity resin prior to elution prevents considering a co-expression mini-test as negative when, in fact, the complex is assembled but requires different purification conditions or slightly modified protein constructs to behave correctly in solution. An alternative to this protocol is to incubate the affinity resin with elution buffer (e.g., with a buffer containing imidazole in case of poly-histidine tag affinity purification), store the elution supernatant for analysis, wash the beads with washing buffer, and finally incubate the beads with Laemmli buffer to check for precipitated samples.
43. Although the presence of several proteins on the same lane of a gel is a good indication of an interaction between these proteins, it is important to keep in mind that some false-positives can be observed. Firstly, in some cases, degradation of the tagged protein might give degradation products that are at the same size as the untagged protein(s). The use of different constructs for the proteins might help resolve this issue. Another problem that might be observed is the strong nonspecific binding of untagged proteins. Expression of the untagged proteins on their own (i.e., without co-expression of the poly-histidine-tagged protein) either during the initial co-expression tests or once an interaction has been apparently observed should help resolve this issue.
44. When the boundaries of the constructs used in co-expression have not been precisely defined for the proteins studied, these

constructs and the complex they form might be present in low quantity in the lysis supernatant, for instance, due to poor expression or poor solubility of one partner, and difficult to distinguish from the *E. coli* proteins bound non-specifically to the affinity resin. Spending time analyzing the gels can be rewarding providing first clues to complex formation and reconstitution.

45. Co-expression experiments can fail because the constructs used for some of the proteins are either too short or too long, leading to lack of interaction in the former case and possible precipitation of the complexes in the latter. The use of many different constructs, as for single protein expression, defined either using bioinformatics analysis or empirical means, is therefore recommended if no interactions are observed. In addition, some proteins may not only require their interaction partners to be soluble but also need to be fused to a specific fusion protein. The use of larger fusion proteins should therefore be considered in a second round of co-expression experiments if the first tests failed.
46. The auto-induction medium and also the medium generally used for co-expression mini-tests (*see* Subheading 2.8) are not optimal for large-scale co-expressions. Use rather 2× LB medium with induction with IPTG (1 mM final) for large cultures.
47. For the concatenation reaction (*see* [8] and http://archive.igbmc.fr/recherche/Sup_papers/Cavarelli/pET-MCN/index.html), the vector that will accommodate the new gene (pET-MCN series) or the new promoter (pET-MCP series) is termed acceptor vector. This acceptor vector already harbors a gene in its own promoter and is cut with the enzymes BglII and SpeI (*see* **Note 11**). The vector that provides the new gene/new promoter is called donor vector and is cut with BglII and XbaI (alternatively NheI or AvrII; *see* **Note 11**). Note that these acceptor and donor vectors are different from the ones used in the ACEMBL system.
48. Shrimp alkaline phosphatase is used to dephosphorylate the acceptor vector to prevent its religation in the subsequent ligation reaction. Shrimp alkaline phosphatase is preferred to other phosphatases since it can be used with a large variety of buffers.
49. On agarose gel, there should be a single band corresponding to the cut acceptor vector, the fragment removed being very small. For this reason it is even possible to skip separation on agarose gel and directly purify the acceptor vector with the NucleoSpin Gel and PCR Cleanup Kit (Macherey-Nagel).
50. For the donor vector, two bands should be present: the cut vector and a smaller band which should have the size of the

gene increased by 100 base pairs (approximate size of the fragment between the T7 promoter and the ribosome-binding site) plus the size in base pairs of the fragment coding for the tag/fusion, if any. The smaller band should be cut away and purified with a NucleoSpin Gel and PCR Cleanup Kit.

51. In some cases, concatenation can lead to unexpected recombination events in the promoter region. Sequencing of the promoter region should be performed to ensure that the plasmids selected did not undergo this kind of recombination. It is recommended that at least two clones are sequenced to ensure that a correct version is obtained.

Acknowledgments

The authors are supported by institutional funds from the Centre National de la Recherche Scientifique (CNRS), the Institut National de la Santé et de la Recherche Médicale (INSERM), the Université de Strasbourg (UDS), the French Infrastructure for Integrated Structural Biology (FRISBI; ANR-10-INSB-05-01), and Instruct, part of the European Strategy Forum of Research Infrastructures (ESFRI), and supported by national member subscriptions. IB acknowledges support from the European Commission (EC) Framework Programme (FP) 7 ComplexINC project (Contract Nr. 279039).

References

1. Kerrigan JJ, Xie Q, Ames RS, Lu Q (2011) Production of protein complexes via co-expression. *Protein Expr Purif* 75:1–14
2. Vincentelli R, Romier C (2013) Expression in *Escherichia coli*: becoming faster and more complex. *Curr Opin Struct Biol* 23:326–334
3. Perrakis A, Romier C (2008) Assembly of protein complexes by coexpression in prokaryotic and eukaryotic hosts: an overview. *Methods Mol Biol* 426:247–256
4. Romier C (2008) Protein complexes assembly by multi-expression in bacterial and eukaryotic hosts. In: Sussman J (ed) *Structural proteomics*. World Scientific Publishing Company, London, pp 233–250
5. Romier C, Ben JM, Albeck S et al (2006) Co-expression of protein complexes in prokaryotic and eukaryotic hosts: experimental procedures, database tracking and case studies. *Acta Crystallogr D Biol Crystallogr* 62:1232–1242
6. An Y, Meresse P, Mas PJ, Hart DJ (2011) CoESPRIT: a library-based construct screening method for identification and expression of soluble protein complexes. *PLoS One* 6:e16261
7. Bieniossek C, Nie Y, Frey D et al (2009) Automated unrestricted multigene recombineering for multiprotein complex production. *Nat Methods* 6:447–450
8. Diebold ML, Fribourg S, Koch M et al (2011) Deciphering correct strategies for multiprotein complex assembly by co-expression: application to complexes as large as the histone octamer. *J Struct Biol* 175:178–188
9. Fribourg S, Romier C, Werten S et al (2001) Dissecting the interaction network of multiprotein complexes by pairwise coexpression of subunits in *E. coli*. *J Mol Biol* 306:363–373

10. Held D, Yaeger K, Novy R (2003) New coexpression vectors for expanded compatibilities in *E. coli*. in *Novations* 18:4–6
11. Novy R, Yaeger K, Held D, Mierendorf R (2002) Coexpression of multiple target proteins in *E. coli*. in *Novations* 15:2–6
12. Scheich C, Kummel D, Soumailakakis D et al (2007) Vectors for co-expression of an unrestricted number of proteins. *Nucleic Acids Res* 35:e43
13. Selleck W, Tan S (2008) Recombinant protein complex expression in *E. coli*. *Curr Protoc Protein Sci*. Chapter 5, Unit 5 21
14. Tan S (2001) A modular polycistronic expression system for overexpressing protein complexes in *Escherichia coli*. *Protein Expr Purif* 21:224–234
15. Tan S, Kern RC, Selleck W (2005) The pST44 polycistronic expression system for producing protein complexes in *Escherichia coli*. *Protein Expr Purif* 40:385–395
16. Tolia NH, Joshua-Tor L (2006) Strategies for protein coexpression in *Escherichia coli*. *Nat Methods* 3:55–64
17. Lariviere L, Plaschka C, Seizl M et al (2012) Structure of the Mediator head module. *Nature* 492:448–451
18. Busso D, Peleg Y, Heidebrecht T et al (2011) Expression of protein complexes using multiple *Escherichia coli* protein co-expression systems: a benchmarking study. *J Struct Biol* 175:159–170
19. Vijayachandran LS, Viola C, Garzoni F et al (2011) Robots, pipelines, polypeptides: enabling multiprotein expression in prokaryotic and eukaryotic cells. *J Struct Biol* 175:198–208
20. Li MZ, Elledge SJ (2007) Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. *Nat Methods* 4:251–256
21. Jeong JY, Yim HS, Ryu JY et al (2012) One-step sequence- and ligation-independent cloning as a rapid and versatile cloning method for functional genomics studies. *Appl Environ Microbiol* 78:5440–5443
22. Studier FW (2005) Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif* 41:207–234

Chapter 5

The Production of Multiprotein Complexes in Insect Cells Using the Baculovirus Expression System

Wassim Abdulrahman, Laura Radu, Frederic Garzoni, Olga Kolesnikova, Kapil Gupta, Judit Osz-Papai, Imre Berger, and Arnaud Poterszman

Abstract

The production of a homogeneous protein sample in sufficient quantities is an essential prerequisite not only for structural investigations but represents also a rate-limiting step for many functional studies. In the cell, a large fraction of eukaryotic proteins exists as large multicomponent assemblies with many subunits, which act in concert to catalyze specific activities. Many of these complexes cannot be obtained from endogenous source material, so recombinant expression and reconstitution are then required to overcome this bottleneck. This chapter describes current strategies and protocols for the efficient production of multiprotein complexes in large quantities and of high quality, using the baculovirus/insect cell expression system.

Key words Recombinant protein complex production, BEVS, Baculovirus, Insect cells, Infection and coinfection methods, Multigene expression, Multiprotein complex

1 Introduction

Most eukaryotic proteins form transiently or stably interlocking assemblies that often contain many subunits. Obtaining these assemblies in a purified form in high quality and quantity is crucial for research aimed at understanding how these protein machines function in a physiological context as well as for drug discovery applications. In the absence of their interacting partners, proteins are often insoluble, improperly folded, or non-functional. Furthermore, protein complex composition and activity may vary depending on tissue type, cell state, and also on modifications of the subunits (e.g., phosphorylation, acetylation, and methylation). Production of multiprotein complexes of higher eukaryotes, including those of human origin, in a suitable form to study their structure and function can be extremely challenging. The low natural

abundance and heterogeneity of native complexes in cells often prevent the extraction and use of endogenous sources for the purification of protein complexes.

Recombinant approaches have become the method of choice for the production of stable macromolecular assemblies. Certain complexes may be reconstituted from recombinant proteins produced separately. Such reconstitution methods are relatively simple and particularly useful when one component of the complex is not a protein (e.g., DNA or RNA) or when the complex is short-lived or transient and cannot be purified intact. In many cases, however, this strategy is not applicable as individual subunits of a complex often cannot be expressed and manipulated in the absence of their natural partners. Multicomponent systems, including not only self-assembling multiprotein complexes but also proteins requiring specific chaperones to assist folding or post-translational modifications crucial for biological activity, will then require coexpression of a number of proteins. Many examples have shown that the simultaneous expression of different subunits of a protein complex facilitates their folding, promotes solubility, and limits degradation of regions that fold upon binding.

A number of coexpression systems are available based on *Escherichia coli* (*E. coli*) (see Chapter 6), mammalian cells (see Chapter 8), and insect cells using baculovirus vectors. Although *E. coli* is robust and inexpensive as a host, there are a number of limitations in using bacteria for synthesis of eukaryotic proteins. In particular, bacteria are unable to provide post-translational modifications and folding aids such as chaperones required for the generation of fully functional eukaryotic proteins [1, 2]. In contrast to *E. coli*, insect cells and mammalian cells have the machinery for proper folding, post-translational modification, authentic processing, and correct targeting of expressed proteins [3–5]. Several developments have been made to the baculovirus expression system, which make this the system of choice for the expression of multiprotein eukaryotic complexes. These include streamlining the assembly of multigene vectors and engineering the baculovirus genome to optimize the expression levels (for example, the MultiBac system [6–8]).

In this chapter, we describe protocols for the production of multiprotein complexes in insect cells using baculovirus expression vector systems. We describe standard procedures for working with the baculovirus system and detail the two main strategies for coexpression of multiple proteins in the same cell, which are (1) coinfection of insect cells with several viruses, each expressing a single protein and (2) infection with one single baculovirus containing all heterologous genes of choice 9.

2 Materials

2.1 Cell Culture

1. Plasticware for monolayer culture of insect cells: 25 cm² tissue culture flasks, six-well tissue plate with flat bottom, low evaporation lid, petri Style tissue culture dish (D60×H15 mm) with 2 mm grid.
2. Plasticware for suspension culture of insect cells: 50 ml polypropylene tube with filter cap for oxygenation, glass or disposable Erlenmeyer flasks in different sizes.
3. Plate sealer, breathable, gas permeable, 80×150 mm.
4. Sf9, Sf21, and High Five cells adapted for suspension growth.
5. TNM-FH and serum-free insect cell medium.
6. Foetal Bovine Serum (FBS).
7. 1.3× SF900 medium for plaque assay.
8. Cell culture grade DMSO.
9. Cell culture grade BSA.
10. Insect cell freezing solution: 90 % insect cell medium, 10 g/L BSA, 10 % (v/v) DMSO. Sterilize solution by filtration.
11. Trypan blue: 0.4 % solution in PBS, pH 7.2.
12. Temperature controlled room or incubator set at 27 °C.
13. Platform for spinner flask operating at 27 °C and stirring up to 150 rpm.
14. Orbital shaker fitted for 250 ml to 2 l Erlenmeyer flasks, with shaking speed of up to 150 rpm (125 mm orbital). For cultures in 50 ml polypropylene tubes shake up to 250 rpm.
15. Inverted phase-contrast microscope or optionally fluorescence microscope.
16. Cell-counting chamber or optionally automated cell counter.
17. Centrifuge with adaptors for 1 l, 250 ml, 50 ml, and 15 ml tubes.

2.2 Molecular Biology

1. Commercial DNA purification kits for small scale and large scale DNA preparation.
2. 3 M Na acetate, pH 5.2.
3. T4 DNA ligase.
4. LB medium, LB agar medium.
5. IPTG, 1 M stock solution in water.
6. X-gal, 100 mg/ml stock solution in dimethylformamide.
7. Bsu36I restriction enzyme.
8. Low salt LB (for 100 ml: 1.0 g Bacto-Tryptone, 0.5 g Bacto-yeast extract, 0.5 g NaCl).

Table 1
Reagents required to generate recombinant viruses

	Tn7 transposition-based system (Method 1) (Bac-to-Bac, Multibac)	Homologous recombination (Method 2)
Method	Transfection of recombinant viral DNA	Co-transfection of linearized viral DNA with a transfer vector
Viral DNA	DH10Bac ^a , DH10MultiBac, DH10EMBacY [6, 7]	BAC10:KO1629 in <i>E. coli</i> DH10B [13]
Transfer vectors (acceptors)	pFastBac ^a pFastBac Dual ^a , pKL, pFL [6] pACEBac1, pACEBac2 [15]	pBacPAK8 ^b , pAC8 [11] pACAB3, PACAB4 [10] pAC8_DsRed ^c , pAC8_MF ^c
Transfer vectors (donors)	pSPL, pUCDM, pIKD, pIDK, pIDC, ... [6, 15]	
Transfer vector (polyprotein)	pPBac, pKL-pBac [12]	

^aInvitrogen™

^bClontech™

^cAvailable on request; (Kolesnikova et al. in prep)

9. L-Arabinose.

10. *E. coli* DH5α and TOP10 strain.

11. pBAD-His-Cre plasmid [7].

12. Tetracycline, 15 mg/ml stock solution in ethanol.

13. Kanamycin, 50 mg/ml stock solution in water.

14. Gentamicin, 10 mg/ml 7 mg/ml stock solution in water.

15. Chloramphenicol, 34 mg/ml stock solution in ethanol.

16. Zeocin, 10 mg/ml stock solution in water.

Reagents to generate recombinant baculovirus are listed in Table 1. Complementary set of oligonucleotides that correspond to the LoxP site:

Fw: 5': ATAACCTTCGTATA GCATACAT TATACGAAGTTAT 3'

Rev: 5': ATAACCTTCGTATA ATGTATGC TATACGAAGTTAT 3'

2.3 Other Buffers

1. Cre purification buffer: 20 mM Tris-HCl (pH 8.0), 300 mM NaCl, 10 % glycerol, 5 mM imidazole, 5 mM DTT.
2. 10× Cre-lox reaction buffer: 0.5 M Tris-HCl (pH 7.5), 0.33 M NaCl, 0.1 M MgCl₂.
3. Lysis buffer: 20 mM Tris-HCl (pH 7.5), 250 mM NaCl, 1 mM DTT, 0.1 % NP40, containing protease inhibitor cocktail (PIC) and type 1 DNAase.

4. Wash buffer: 20 mM Tris-HCl (pH 7.5), 250 mM NaCl, 1 mM DTT, 0.1 % NP40.
5. Elution buffer: 20 mM Tris-HCl (pH 7.5), 250 mM NaCl, 1 mM DTT, 0.1 % NP40 with appropriate elution agent.

3 Methods

The simultaneous production of several proteins in insect cells using the BEVS requires the delivery of various genes, either by a number of individual baculoviruses (coinfection, Fig. 1a–c) or by employing a single virus that contains several genes (multigene virus, Fig. 1d–f). Coinfection enables exploratory screening of

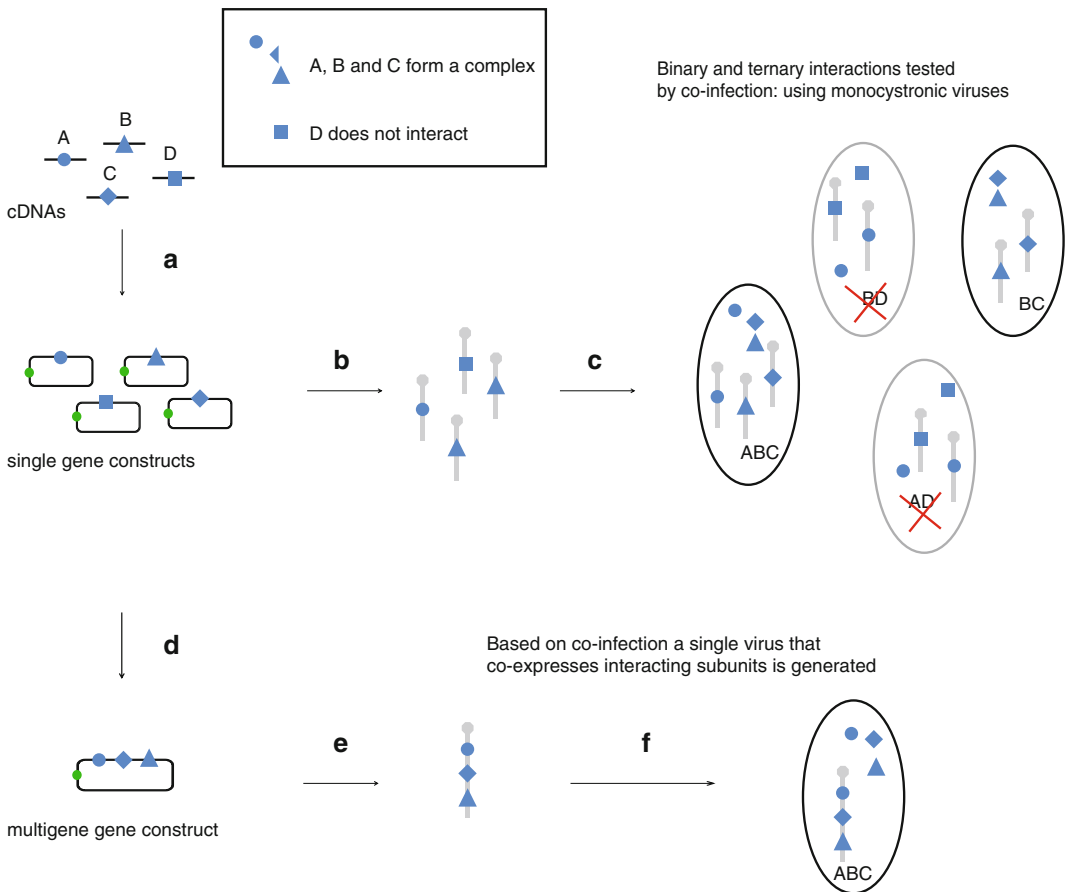


Fig. 1 Strategy for reconstitution of multiprotein complexes: **(a)** cDNAs of potential targets are cloned into transfer vectors and **(b)** the corresponding single-gene baculoviruses (expressing subunits A, B, C, and D, for example) are generated. **(c)** Insect cells are coinfecting by two or three single-gene baculoviruses and association between subunits are tested by pull down, for example. **(d)** Genes encoding proteins that form stable complex (subunits A, B, and C in this example) are assembled in a single multigene transfer vector **(e)** to generate a recombinant multigene baculovirus. **(f)** Insect cells infected by multigene baculoviruses express the identified multiprotein complex A–B–C

putative interaction partners prior to large-scale expression but necessitates the maintenance of many viruses at known titers. The use of multigene baculoviruses ensures that all proteins necessary for the formation of the recombinant complex are expressed in the same infected cell, which greatly simplifies the management of the experiment.

Coexpression of multiple genes often requires extensive screening efforts to identify suitable constructs: mutations and/or deletions to optimize expression and solubility, nature and position of the affinity tag to facilitate purification. In the absence of prior knowledge about the proteins of interest, a first set of experiments with viruses expressing a single protein will provide valuable information on expression level/solubility of individual subunits. At this stage, different affinity tags can be tested [11]. For routine productions and for production of multiprotein complexes composed of a larger number of subunits, it is however preferable to use a single virus that coexpresses the different proteins. To obtain this virus, based on initial experiments, validated single gene expression units should be assembled into multigene expression cassettes using restriction/ligation or sequence and ligation independent cloning (SLIC) methods. Cre-mediated fusion of acceptor vectors with specific donor plasmids as well as the use of polyproteins for balancing subunit stoichiometry offer further options to efficiently coexpress a large number of genes [7, 8, 15].

3.1 Insect Cell Management

Working with baculoviruses requires a basic knowledge of general cell culture methods and insect cell physiology. Insect cells and viruses are handled in a laminar flow hood under aseptic conditions preferentially in absence of antibiotics, as these can mask low levels of contamination (*see Note 1*). All cell culture experiments are carried out at 27 °C, either in incubators or ideally in a room conditioned at this temperature. Doubling rate of the majority of insect cells (i.e., the time while the number of cells per volume increases by a factor of 2) at this temperature is around 18–20 h. The density of insect cells should range between 0.5 and 2.0×10^6 cells/ml, especially during expression experiments.

1. Remove vial of cells from liquid nitrogen and place in water bath at 37 °C. Thaw rapidly with gentle agitation until cells are almost thawed and remove cells from the water bath. Leaving cells at 37 °C after they have thawed will result in cell death.
2. Decontaminate the outside of the vial by spraying with 70 % ethanol, dry and place on ice.
3. Pre-wet a 25 cm² tissue culture flask by coating the adherent surface with 4 ml media.
4. Transfer the 1 ml thawed cell suspension directly into the 4 ml of media.

5. Incubate flask at 27 °C and allow cells to attach for 30–45 min.
6. After cells are attached, gently remove the medium (as soon as possible to remove the freezing solution containing DMSO).
7. Feed cells with 5 ml of fresh medium.
8. After 24 h, change to fresh medium.
9. Allow cells to grow until 90 % confluence. Detach cells by tapping the flask or by sloughing (streaming medium over the monolayer with a pipette to dislodge cells) and initiate suspension culture (Subheading 3.1, step 2).
10. Take an aliquot of the cell suspension (Subheading 3.1, step 1), count cells and determine their viability (*see* Note 2).
11. Add an appropriate volume of medium to a sterile Erlenmeyer, inoculate with cells to obtain a starting density of 0.5×10^6 cells/ml and incubate cells at 27 °C with agitation (80–100 rpm).
12. Monitor culture daily until cell density reaches $2\text{--}3 \times 10^6$ cells/ml and seed a fresh Erlenmeyer as in step 2.
13. Count cells and ensure that you have enough cells for preparing 2–4 vials (Table 2).
14. Prepare cryovials, cool them on ice.
15. Centrifuge cells at $100\text{--}150 \times g$ for 10 min at RT and remove supernatant. If High Five cells are used, make sure to keep the conditioned media when preparing freezing media.
16. Resuspend cells at the density indicated in Table 2 in the proper media.
17. Transfer 1 ml to each sterile cryovial.
18. Place at -20 °C for 1 h, then store at -80 °C for 24–48 h.
19. Transfer to dewars filled with liquid nitrogen for long-term storage.

Table 2
Media composition for freezing most commonly used cell lines

Cell line	Freezing media	Cell density (cells/ml)
Sf21, Sf9	60 % Sf900 medium (GIBCO) 30 % Fetal Bovine Serum (FBS) 10 % DMSO	1×10^7
High Five	42.5 % conditioned Express5 medium 42.5 % fresh Express5 medium 5 % FBS 10 % DMSO	3×10^6

3.2 Generation of Recombinant Baculoviruses

Two methods are available for the generation of recombinant baculoviruses; both make use of the baculovirus genome engineered into a bacmid for propagation in *E. coli*. The first method is based on site-specific transposition (Tn7 transposition) of an expression cassette into the baculovirus genome in *E. coli*. The second employs a transfer vector and viral DNA that are cotransfected into insect cells and utilize host enzyme-mediated homologous recombination.

3.2.1 Method 1: Tn7 Transposition and Preparation of Bacmid for Transfection

1. *Day 1*: transform 50–100 μ l chemical-competent cell solution (DH10Bac, DH10MultiBac) with 10–100 ng of appropriate transfer vector, resuspend in 600 μ l of LB (or SOC as preferred), and incubate cell solution at 37 °C overnight (8 h).
2. *Day 2*: streak out 150 μ l on selection plates for blue/white screening containing the relevant antibiotics (Kanamycin at 50 μ g/ml, Tetracyclin at 10 μ g/ml, and Gentamicin at 7 μ g/ml), IPTG (1 mM), and X-gal (100 μ g/ml). Incubate plates at 37 °C until blue or white colored colonies can be unambiguously seen. Use dilution series (1:1, 1:10, 1:100, 1:1,000) in order to obtain optimal separation of colonies on one of the plates.
3. *Day 3*: restreak several white colonies and one blue colony as a control on the same plate in **step 2** for each construct to confirm the color of colonies.
4. *Day 4*: pick two white colonies for each construct, start 2 ml minicultures (in LB with appropriate antibiotics) overnight at 37 °C. Note that pellets can be, at this step, frozen at –20 °C for short-term storage, or kept as glycerol stock. Avoid long-term of purified bacmid (at 4 °C or –20 °C).
5. Prepare bacmid DNA using a standard plasmid purification kit, taking care not to vortex the sample during the DNA preparation to avoid shearing of the bacmid DNA. At this stage, the bacmid preparation can be checked on an agarose gel and transposition analyzed by PCR and/or sequencing (*see Note 3*). The recombinant bacmid is now ready for transfection into insect cells to generate the baculovirus.

3.2.2 Method 2: Preparation of Linearized Viral DNA and Transfer Vector for Cotransfection

1. *Day 1*: inoculate 50 ml of LB Broth medium starter culture supplemented with kanamycin at 50 μ g/ml and chloramphenicol at 25 μ g/ml with a glycerol stock of Bac10:KO1629 [13] (*see Note 4*). At the end of the day, use this culture to inoculate 1 l of LB Broth medium supplemented with kanamycin and chloramphenicol, and let the culture grow overnight at 37 °C.
2. *Day 2*: purify the bacmid using a large-scale DNA purification kit (Maxi prep). Follow manufacturer's protocol until DNA

elution and isopropanol precipitation. DNA pellet is then washed with 10 ml of 70 % ethanol.

3. Remove ethanol and air-dry the precipitated bacmid under the sterile hood. Resuspend DNA in 200 μ l of sterile ultrapure water.
4. Transfer 10 μ l of suspension to sterile 1.5 ml tube to estimate DNA concentration and to analyze. We usually obtain 20–40 μ g of purified bacmid from 1 l of *E. coli* culture.
5. Linearize the bacmid with the restriction enzyme Bsu36I. For the digestion of 20 μ g bacmid, mix 200 μ l of bacmid (100 μ g/mL), 30 μ l of 10 \times NEB3 buffer, 3 μ l of 100 \times BSA (optional), 57 μ l of ultrapure water, and 10 μ l of Bsu36I (NEB) (20 U/ μ l).
6. Incubate for 5 h at 37 °C, then analyze an aliquot by gel electrophoresis on a 0.8 % agarose gel to check digestion before heat inactivation (20 min at 72 °C) of Bsu36I.
7. The linearized bacmid can be stored at 4 °C for 3–6 months. Alternatively, prepare aliquots of 6.5 μ g (65 μ l); each aliquot is sufficient for 6–12 small-scale transfections (six-well plate format) and freeze at –20 °C. Once an aliquot was thawed, keep DNA at 4 °C and do not refreeze.
8. Perform a small or medium scale DNA preparation of a suitable transfer plasmid and precipitate 10 μ g of DNA with 300 mM Na-acetate pH 5.2 (final concentration) and three volumes of ethanol 100 %. Place at –80 °C for more than 1 h and centrifuge at 25,000 $\times g$ for 15 min. Carefully remove the supernatant, add 1 ml of cold 70 % ethanol, and centrifuge it again.
9. Remove ethanol and air-dry the precipitated DNA under the sterile hood. Resuspend DNA in 20 μ l of sterile ultrapure water. Take an aliquot to measure the DNA concentration and store at –20 °C.

3.2.3 Transfection of Insect Cells to Generate Recombinant Baculovirus

1. Seed 0.5–1 $\times 10^6$ cells from a stock culture into the wells of six-well tissue culture plates. Add medium in each well to a total volume of 3 ml. In a typical experiment, include one well that contains only cells, one well that contains only medium as well as a positive fluorescent transfection control (PC), e.g., with a DsRed expression plasmid and bacmid. For each DNA construct, seed two wells.
2. For each construct, prepare transfection complexes by diluting the purified bacmid (method 1, Subheading 3.2.1) or a mixture of 0.5–1.0 μ g of linearized viral DNA and 2 μ g of transfer plasmid (method 2, Subheading 3.2.2) in 200 μ l insect cell medium.
3. Mix 100 μ l of insect cell medium with 2–10 μ l of transfection reagent (e.g., Cellfectin II Life technologies) in a separate Eppendorf tube (*see* **Note 5**).

4. Add the diluted transfection reagent to the DNA solution (respect the addition order), vortex for 10 s, and incubate at room temperature (RT) for 20 min.
5. Add 1 ml of medium to the transfectant-DNA suspension and use it to replace supernatant from seeded cells. Incubate for 5 h at 27 °C, aspirate the suspension, add 3 ml of fresh medium, and return to 27 °C.
6. Incubate for a maximum of 5–6 days at 27 °C. When more than 50 % of cells in the positive control express the fluorescent protein, carefully collect the supernatant and store it at 4 °C, protected from light. This is the passage 0 (P0) virus stock. The P0 stock can be used directly to test expression of the desired proteins in small-scale experiments or has to be concomitantly amplified for production.

3.3 Virus Amplification and Storage

3.3.1 Amplification

1. Prepare 100 ml Erlenmeyer flask containing 25 ml of *Sf9* or *Sf21* suspension at a density of 0.5×10^6 cells/ml in exponential growth phase, infect with a small volume of P0 virus, and incubate at 27 °C with agitation (100 rpm). The volume of virus depends on the titer of your virus stock and thus on transfection/cotransfection efficiency (*see* **Note 6**).
2. Count cells 1 day postinfection and measure their size distribution. If the volume of virus added to the culture was adequate (MOI of 0.5 or below), cells should look healthy and should have doubled. At this time, infected cells should be releasing budded virus into the medium to infect other cells. If too much virus was added, you should see signs of infection: cells swell (size can increase up to 20–30 %), stop dividing, and appear uniformly rounded with enlarged nuclei. Restart with less virus.
3. Count cells 2 days postinfection and measure their size distribution. Most, if not all, of the cells should show substantial swelling, and proliferation arrest should be observed. The cell count number should be substantially below that expected if they had continued doubling every 24 h. Return the flask to a 27 °C shaker for 24 h.
4. Count cells again 24 h later and estimate their size.
5. If substantial swelling and proliferation arrest was already observed the day before, there should be no increase in cell number. Go to **step 7**.
6. If not, return the flask to a 27 °C shaker for another 24 h until proliferation arrest is determined. Dilute culture (and split in fresh flask if necessary) to maintain cell density below 2×10^6 cells/ml to prevent oxygen deprivation and entry into stationary phase. If cells do not stop doubling after 5 days, we recommend repeating the initial virus preparation.

7. The culture should be harvested when cells have been infected for about 48 h, i.e., 24 h after cells have stopped dividing. Transfer the suspension into a fresh 50 ml tube and spin for 5 min at $4,000\times g$ in a table top centrifuge. Collect the supernatant into a fresh 50 mL tube and supplement with 10 % FBS if a serum-free medium was used and store at 4 °C protected from light. This is the P1 virus stock.
8. Gently resuspend the pellet in fresh medium (respecting the cell density) and place back suspension into the shaker flask for further analysis.
9. P1 is sufficient for initial protein expression studies. If large volumes of virus are required, repeat the cotransfection or amplify P1 to obtain P2 (and eventually P3) (*see Note 7*).
10. Detailed monitoring of virus amplification is not always possible when a large number of viruses are needed simultaneously. Simplified protocols are described below:

Alternative protocol 1 (5–7 days): Infect a 250 ml suspension culture in exponential growth phase at $0.5\text{--}1\times 10^6$ cells/ml by adding 0.25 ml of P0 stock and incubate under agitation at 27 °C. After 5–7 days, observe cells for signs of infection, spin the culture, and harvest the supernatant (P1 stock).

Alternative protocol 2 (3 days+3 days): Infect a 25 ml suspension culture in exponential growth phase at $0.5\text{--}1\times 10^6$ cells/ml by adding 2.0 ml of P0 stock and incubate with agitation at 27 °C for 3 days. Spin the culture and harvest the supernatant (P1 stock).

Infect a 250 ml suspension culture in exponential growth phase at $0.5\text{--}1\times 10^6$ cells/ml by adding 2.5 ml of P1 stock and incubate under agitation at 27 °C for 3 days. Spin the culture and harvest the supernatant (P2 stock).

3.3.2 Storing Viruses (BIIC)

Safe storage of valuable virus stocks for future use is often of paramount importance. Virus stocks can be stored at 4 °C in the dark (after addition of 10 % FBS or 0.1–1 % BSA) for 1–12 months. We found that storing virus at liquid nitrogen temperatures in frozen aliquots of baculovirus-infected insect cells (BIICs) is most advantageous in terms of virus stability and storage space requirements. Frozen aliquots of infected insect cells can be prepared and rethawed and used for protein expression without loss after extended storage times (several years). Integrity of recombinant virus can be checked by PCR (*see Note 8*).

1. Prepare 50 ml culture of *Sf9* or *Sf21* cells in exponential growth phase at 1×10^6 cells/ml.
2. Infect cells with a chosen volume of P1 virus.

3. Maintain cells to a concentration of 1×10^6 cells/ml until proliferation arrest.
4. Centrifuge cell culture in sterile 50 ml tube at $100\text{--}150 \times g$ for 10 min, remove supernatant.
5. Resuspend cells gently in sterile freezing solution to a final density of 1×10^7 cells/ml.
6. Transfer 1 ml aliquots into sterile cryovials.
7. Place at -20°C for 1 h.
8. Store at -80°C for 24–48 h.
9. Store cryovials in liquid nitrogen for indefinite time.

3.4 Protein–Protein Interaction Screening by Coinfection

A set of viruses for expression of all components of the multiprotein complex are generated as described above. In the case of large subunits, this could also include subdomain constructs designed from the analysis of multiple sequence alignments, predictions of secondary structure, and disordered regions. Sequence tags for detection/purification are incorporated at either the amino terminus or carboxy terminus of the constructs, e.g., His-tag, StrepII tag. Using different tags for each component enables the expression of each to be followed, though in practice depending on the number of subunits in the complex, this may be limited by the availability of tags.

3.4.1 Construction Transfer Vectors Expressing Individual Subunits

1. Compile the cDNAs encoding the subunit sequences of interest and select a set of transfer vectors (Fig. 2a) that will be used to generate the first ensemble of recombinant viruses. Bear in mind that in a second set of experiments, you might have to generate multigene expression constructs and use an adapted transfer vector harboring a multiplication module and/or a LoxP site from the MultiBac suite to facilitate further DNA manipulations (*see* Subheading 3.5).
2. Design PCR primers for insertion of the genes into the selected transfer vectors using either ligation independent cloning technology (e.g., SLIC) or conventional restriction enzyme/ligation based method. If the transfer vectors do not encode the desired affinity/epitope tags, use nested or assembly PCR to incorporate the required nucleotide sequences.
3. For ligation independent cloning, we recommend to design primers with 20–30 bp of homology to the destination vector and to linearize the acceptor vector with two restriction enzymes (for *in silico* design web tools, *see* Note 9).
4. Generate recombinant baculoviruses and amplify to 100 ml stock as described in Subheading 3.3.

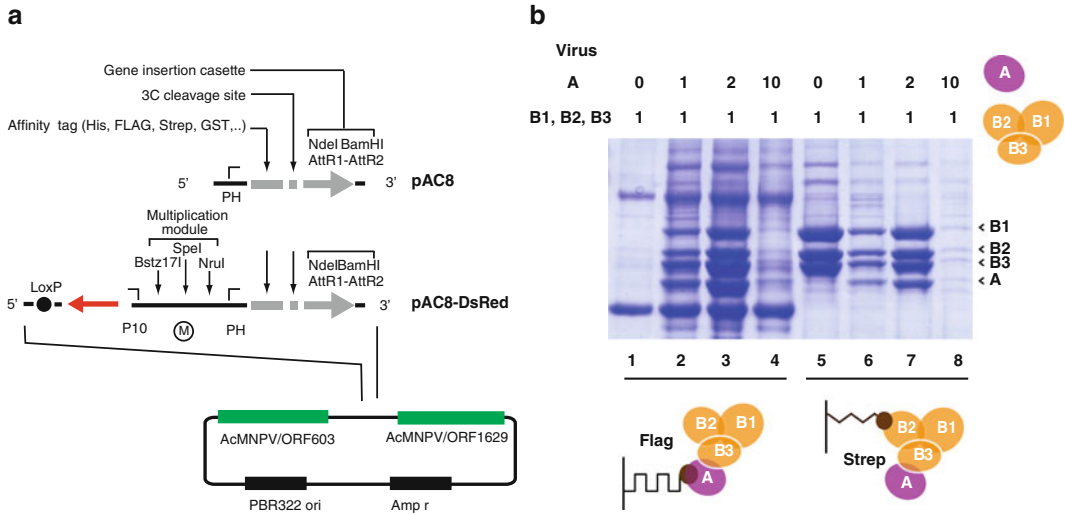


Fig. 2 Interaction screening. **(a)** pAC8 baculovirus expression shuttle vectors are tailored for interaction screening and multiprotein complex production in insect cells. This set is derived from pBacPAK8 (BD Biosciences), a shuttle vector that is used to obtain recombinant virus by homologous recombination in insect cells (homology regions are in *green*) and express the gene of interest under the dependence of the polyhedrin promoter (PH). The tag sequence is followed by a precision protease (3C) cleavage site and a gene insertion cassette (pair of unique restriction sites (NdeI and BamHI) or a Gateway cloning cassette) [9]. pAC8-DsRed derivatives also harbor a DsRed expression cassette placed under control of P10 promoter to monitor infection, a multiplication module (M) to insert an additional expression cassette as well as a loxP sequence for Cre-mediated vector fusion. **(b)** Insect cells were coinfecting with baculovirus A (for expression of protein A fused with an N-terminal Flag tag) and baculovirus B1–B2–B3 (for expression of proteins B1, B2, and B3; B2 harbors strep-tag). Cells were infected with the same amount of virus A (corresponding to an MOI of 1) and increasing volumes of virus B1–B2–B3 (MOIs of 1, 2, and 10). After cell lysis, equal parts of clarified extract were mixed with anti-flag M2 beads (lanes 1–4) or streptactin-coated beads (lanes 5–8). After extensive washing, SDS loading dye was added to the beads, and samples were analyzed by SDS-PAGE. Anti-flag light and heavy chains are indicated by *black spheres*. Proteins are indicated by *arrowheads*

3.4.2 Small-Scale Expression Test

Small-scale expression tests are carried out to optimize protein production for each component of a complex individually (if expressed). Key parameters are the amount of virus and the time of infection: (1) when infecting cells for protein production, the objective is to get all cells infected synchronously. Typically, conditions that correspond to MOIs in the range of 0.5–10 are tested. (2) The best time to harvest depends on the nature of the target protein. Cells are usually analyzed 48, 72, and 96 h postinfection. Some stable proteins might accumulate to high levels 72 or 96 h postinfection while others, sensitive to degradation, will need to be collected after 24 h or most commonly 48 h. Protein expression may also depend on cell type, so expression using *Sf9*, *Sf21*, or High Five cells has to be tested.

Next, physical interactions between proteins are mapped via coinfection with multiple baculoviruses. Virus stocks from two or more viruses, each expressing a single protein (or a defined set of

proteins known to interact), are used to coinfect insect cells and thus coexpress the proteins of interests. Extracts are subjected to small-scale affinity purifications to identify protein pairs/complexes that tightly interact. There are no generic takings to determine the optimum ratio of viruses for coinfection. Optimized MOIs determined for each individual protein (Subheading 3.4.2) are a good starting point for coinfection but these conditions need to be reoptimized and several MOI combinations have to be tested. Special attention should be given to the fact that varying the MOI ratio of infecting viruses has major impact on protein expression of individual subunits and that unfavorable settings can lead to a significant decrease of the global protein production yield (Fig. 2b) see 15.

We describe a protocol for small-scale optimization of protein expression in suspension cultures in 125 ml Erlenmeyer flasks by infection with one or more baculoviruses. Optimized conditions can be scaled up to 2 and 5 l flasks for medium/large scale productions can be compared.

1. Dispense 1×10^6 cells (*Sf9*, *Sf21*, or High Five) from a culture in exponential growth phase into polypropylene tubes equipped with a filter cap for oxygen supply.
2. Pellet cells by centrifugation at $200 \times g$ for 10 min and discard the supernatant.
3. Add the desired volume of each viral stock to the cell pellet and incubate at 27 °C for 1 h with gentle agitation. Different ratios can be tested for each virus (Fig. 2b).
4. Resuspend cells in fresh medium at 1×10^6 cells/ml and incubate at 27 °C for either 48 or 72 h after infection.
5. Centrifuge the cell suspension at $200 \times g$ for 10 min in 50 ml tubes, resuspend cells in 3 ml PBS + 10 % glycerol, centrifuge again, and store pellets at -80 °C.
6. Resuspend cells in 0.8–1.5 ml lysis buffer and break cells by sonication (*see Note 10*). Collect 15 µl aliquots and add 5 µl of 4× SDS loading dye (total extract).
7. Clarify the lysate by centrifugation at $6,500 \times g$ for 60 min at 4 °C and optionally filter the supernatant using a 0.2 µm filter plate. Take a 15 µl aliquot and add 5 µl of 4× SDS loading dye (soluble extract).
8. Incubate the soluble extract with equilibrated affinity resin at 4 °C. Use 25 µl of resin for batch purification and incubate for 15–120 min with slow end-over-end mixing. To facilitate interpretations, flag at least two different proteins with two different tags and split the extract for parallel purification (Fig. 2). Use 100 µl for spin-column or filter-based chromatography. Extended incubation is not recommended as it exposes the sample to protein degradation.

9. Wash the resin with lysis buffer without PIC and elute with 50 μ l of appropriate elution buffer for batch purification or with 200 μ l of elution buffer for spin-column or filter-based chromatography. Take a 15 μ l aliquot from each elution and add 5 μ l of 4 \times SDS loading dye for analysis (eluted fraction).
10. As an alternative to **step 4**, add 25 μ l SDS loading dye to the beads and boil the sample during 2 min prior to SDS-PAGE analysis.

At this stage, if a suitable complex is identified, production can be optimized and/or scaled up. MOIs and virus ratios as well as cell densities at infection influence the necessary duration of the culture and deserve careful analysis. Protein expression/stability may also be affected by the cell line and expression obtained using Sf9, Sf21, or High Five cells can be compared. Storage of frozen BIIIC stocks will facilitate scale-up and development of reproducible production process (Subheading 3.6.2).

3.5 Construction of Multigene Baculoviruses

A number of baculovirus transfer vectors are available to enable the generation of multigene baculoviruses. These transfer vectors include pAcAB3, pAcAB4, pAcUW51, and pFastBacDual that contains two expression cassettes (Table 1). In pFastBacDUAL, PCR fragments encoding proteins of interest can be cloned into expression cassettes driven by either the p10 or polyhedrin promoters using any of the unique restriction endonuclease sites located immediately downstream of the promoters. The MultiBac technology (Fig. 3) has introduced new transfer vectors with enhanced capabilities for DNA manipulations using a multiplication module for concatenating expression cassettes, and LoxP sites for fusing Donor and Acceptor plasmids, each containing one or several expression cassettes, by means of in vitro Cre-LoxP reaction catalyzed by Cre recombinase [6–8].

3.5.1 Modification of an Acceptor Vector

Multigene expression in insect cells using the MultiBac technology relies on tandem recombineering by SLIC and/or Cre-loxP reactions between acceptor and donor plasmids [6, 7]. Any transfer vector can be converted into an acceptor plasmid by insertion of a LoxP sequence between the two DNA elements required for transposition (Tn7R and Tn7L sequence) or for homologous recombination (Orf1629 and lef2/603 sequences). Here we describe how this modification can be achieved for the transfer vector pBacPAK8 (commercially available from Clontech) by insertion of a double-stranded oligonucleotide encoding the LoxP sequence into the unique EcoRV site. Alternatively, if a unique restriction site is not available, PCR-based approaches (such as Quikchange (Stratagene)) can be used as well.

1. Analyze the plasmid map of the transfer vector and decide where the LoxP sequence has to be inserted. Identify a unique

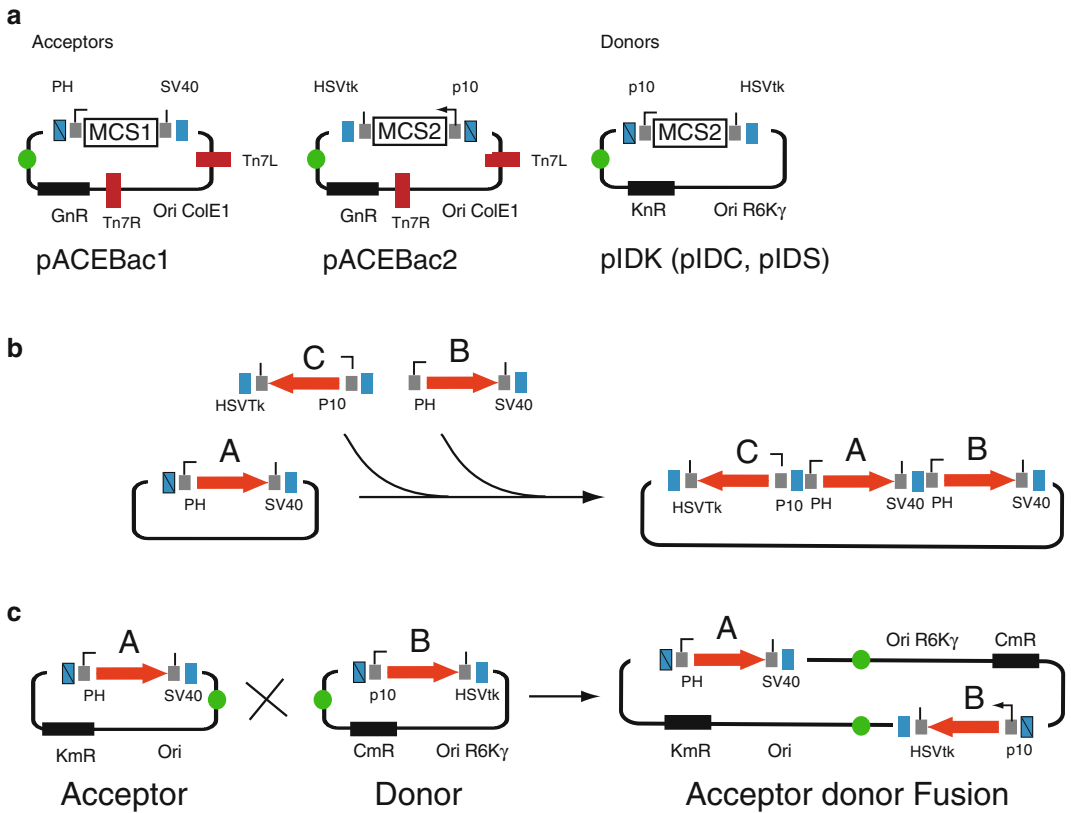


Fig. 3 MultiBac system for generation of multigene baculoviruses [6, 7, 8]. **(a)** MultiBac donor and acceptor plasmids contain multiple cloning sites (MCS) allowing insertion of your gene of interest under the control of late baculoviral promoters (PH or p10) as well as strong eukaryotic polyadenylation signal (from SV40 or HSVtk). All plasmids contain the LoxP sequence (green filled circle) for fusion of donors to an acceptor using Cre-mediated recombination. Acceptors have a regular origin of replication (ori ColE1); whereas, donors have a conditional origin derived from R6K γ phage (ori R6K γ), which allows plasmid replication exclusively in Pir-1 type *E. coli* strains. Each plasmid has a different resistance marker: gentamicin resistance (GnR) for acceptors ACEBac1 and ACEBac2, and kanamycin (KnR), chloramphenicol, or spectinomycin resistance for donors. Multiplication modules facilitating the assembly of several expression cassettes are shown as blue boxes flanking promoter and terminator. Acceptor plasmids contain the DNA sequences (Tn7L and Tn7R) required for transposition by the Tn7 transposase. **(b)** To assemble multigene construct using multiplication module, individual expression cassettes are excised by digestion with a pair of endonucleases and inserted via compatible restriction sites into the multiplication module of a progenitor plasmid. Following the ligation, the restriction sites used for integration are eliminated and multiplication can be repeated iteratively using multiplication module in the inserted cassette. **(c)** Acceptor and donor plasmids contain loxP sequence. Multigene constructs are assembled using Cre-mediated recombination. Resulting multigene plasmid can be propagated in a standard DH5 α strain on the selective medium containing only the antibiotic resistance to which was provided by the donor vector

restriction site located within the DNA fragment that will be transferred into the baculoviral genome by transposition or homologous recombination. Do not use restriction sites located between promoters and terminators as they would interrupt expression units.

2. Digest 5 µg of plasmid with the selected restriction enzyme (EcoRV in our case), treat the plasmid with a phosphatase, isolate the linearized vector from the rest of the reaction using any purification kit available, and quantify.
3. Mix equal volume of 5' phosphorylated complementary oligonucleotides corresponding to the LoxP site (Fwd ATA-ACTTCGTATAGCATACATTATACGAAGTTAT, Rev ATA-ACTTCGTATAATGTATGCTATACGAAGTTAT) at a final concentration of 10 µM in the presence of 5 mM MgCl₂. Heat to 90 °C for 2 min, ramp-cool to room temperature over a period of 45 min, and store at 4 °C.
4. Ligate 10 ng of the double-stranded LoxP fragment and 50 ng of the digested plasmid with T4 DNA ligase according to the manufacturer's protocol. Do not forget to add a negative control without LoxP fragment to evaluate the background from uncut or self-ligating plasmid.
5. Transform the ligation reaction into competent cells and plate onto LB agar containing the appropriate antibiotic.
6. Isolate colonies and purify plasmid DNA. Insertion can be checked by sequencing using a forward oligonucleotide located 150 nucleotides upstream the insertion site (*see* **Note 11**).
7. Perform an in vitro Cre-mediated fusion with a control donor vector expressing a fluorescent protein and generate the corresponding virus.

3.5.2 Cre-LoxP-Mediated Vector Fusion

A transfer vector with a LoxP site can be fused with any other plasmid that also contains a LoxP site using in vitro Cre-mediated recombination. In the MultiBac system, plasmids are divided into donors and acceptors (Fig. 3c). Acceptors contain in addition to the LoxP site, DNA elements for Tn7 transposition into bacmid-based baculovirus genomes. Donor plasmids also contain a LoxP sequence, but in contrast to the acceptor plasmids they contain a conditional R6Kγ replication origin active in *pirI* strains but not in *pir*-negative bacteria such as any commonly used cloning strains (DH5α, Top10, and others). Cre recombinase protein is commercially available but the recombinant protein can be expressed and easily purified from *E. coli* (Subheading 3.5.3).

We provide below a detailed protocol to combine an acceptor with donor(s) (Table 1).

1. Select the donor that will be used and insert the cDNA(s) of interest using your favorite technology (SLIC, RF-cloning, conventional). Don't forget to use a *pirI* strain for cloning and amplification of the plasmid.
2. Mix an equimolar amount of an acceptor and a donor plasmid (250–500 ng) with 1 µl of 10× Cre buffer and adjust the volume to 9 µl with water. Add 1 µl of Cre recombinase, mix briefly, and incubate the reaction for 30 min at 37 °C.

Avoid prolonged incubation that may lead to formation of large plasmids resulting from sequential concatenation reactions. Stop the reaction by heating at 70 °C for 15 min.

3. Transform *pir*-negative chemically competent cells (DH5 α or TOP10, for example) with the Cre-treated donor–acceptor mixture. After 1 h recovery at 37 °C in absence of selection, plate cells on LB agar plates containing the antibiotic(s) to select for the presence of the donor. Selection with antibiotic corresponding to acceptor vector is not required as the propagation of the acceptor–donor fusion depends on the acceptor, which contains a common origin of replication.
4. Select 4–8 colonies. Grow 2 ml cultures in LB supplemented with appropriate antibiotics and isolate plasmid DNA using standard procedures.
5. Verify the recombination by restriction or PCR (*see Note 12*). Optimally, choose restriction enzyme(s) that cut in the acceptor vector and another enzyme that cuts in the vector you are inserting. This strategy will allow you to exclude multiple fusion vectors. Web tools can be used for *in silico* simulation of the Cre reaction (Cre-ACEMBLER) (*see Note 13*).

3.5.3 Purification of Cre Recombinase

Purified Cre recombinase can be purchased from a number of sources. Several bacterial expression plasmids are available and can be used for the production of this enzyme. Here we describe expression in *E. coli* of oligohistidine tagged Cre recombinase under the control of the arabinose promoter [7].

1. Transform DH5 α or Top10 *E. coli* strain with pBADZ-His-Cre plasmid, plate onto low salt LB plates containing 25 μ g/ml of Zeocin (activity of Zeocin depends on ionic strength), and incubate overnight at 37 °C.
2. Prepare a preculture in low salt LB medium containing 25 μ g/ml of Zeocin that will be used to inoculate 100 mL of the same medium. The starting OD_{600nm} should be in the range 0.05–0.1. Place it in the incubator for growth.
3. When the OD_{600nm} reaches 0.5–0.6, add 1 ml of 20 % L-ARABINOSE (0.2 % final concentration) to induce the expression of the recombinant protein. Don't forget to take a 100 μ l aliquot as control before induction (uninduced sample).
4. Let the culture grow for 4 h more, take a 100 μ l aliquot as control (induced sample), and harvest cells by centrifugation at 4,000 $\times g$ for 15 min. Wash the cell pellet with PBS buffer supplemented with 10 % glycerol, centrifuge again, and store the pellet at –20 °C.
5. Add 20 μ l 4 \times SDS loading dye to uninduced and induced sample and boil it for 5 min.

6. Check the expression on 15 % SDS gel, load 10 μ l of the sample without induction and 5 μ l of the sample after induction. The band corresponding to the recombinant Cre protein should be visible at 38 kDa.
7. Resuspend the stored cell pellet in 15 ml of lysis buffer (20 mM Tris-HCl (pH 8.0), 300 mM NaCl, 10 % glycerol, 5 mM imidazole, 5 mM DTT) and sonicate for 1.5 min at 40 % amplitude.
8. Centrifuge at $20,000\times g$ for 20 min at 4 °C. Keep the supernatant.
9. Wash 1 ml IMAC resin with ultrapure water then equilibrate with the lysis buffer. Always handle samples on ice or work in a cold room.
10. Add the equilibrated resin to the supernatant and incubate for 2 h at 4 °C.
11. Centrifuge at $200\times g$ for 10 min to pellet the resin. Discard the supernatant.
12. Wash the resin with 10 ml of lysis buffer containing 10 mM imidazole.
13. Centrifuge again at $200\times g$ for 10 min to pellet the resin and discard the supernatant. Repeat the washing step once more.
14. Poor the resin in an empty gravity flow column and elute with lysis buffer containing 150 mM imidazole.
15. Analyze eluted protein by 15 % of SDS gel, pool the protein containing fractions and dialyze overnight in a buffer containing 20 mM Tris-HCl (pH 8.0), 300 mM NaCl, and 20 % glycerol.
16. Measure the protein concentration and snap freeze 100 μ l aliquots at 0.3 mg/ml concentration. Store at -80 °C.
17. Perform a set of in vitro Cre-mediated fusion reactions with a control vectors using increasing amounts of recombinase for quality control.
18. From 100 ml culture, you should obtain enough recombinase for 100 reactions.

3.6 Scale-Up of Protein Expression

Routine preparation of protein complexes in quantity and quality sufficient for biotechnology applications may necessitate significant efforts, in particular when large volumes of culture are required. Simple protocols to streamline large/medium-scale production are described below.

3.6.1 Use of Fluorescent Reporter Protein

Fluorescent proteins such as the yellow fluorescent protein (YFP) or the *Discosoma* red fluorescent protein (DsRed) can be efficiently used to monitor virus performance and determine the optimal time for harvesting the culture and proceeding to protein purification.

The genes coding for fluorescent proteins can be either fused as a tag to the recombinant protein of choice or integrated directly into the baculoviral genome. In the latter case, if the gene encoding the fluorescent protein is placed under the control of the same promoter that drive the expression of heterologous proteins of interest, detection of fluorescent signal might sign the concomitant expression of proteins of interest. Thus, a plateau of fluorescence is observed when production of heterologous protein is maximal. Toward this goal, the YFP gene was inserted under the control of the polyhedrin promoter into the backbone of the MultiBac virus, giving rise to the virus EMBacY [15].

To determine the optimal time for harvesting cells when protein production is maximal, fluorescence measurements of a reporter gene can be conveniently used as follows:

1. Count cells in an expression culture and withdraw 2×10^6 cells. Centrifuge for 1 min at $15,000 \times g$, remove supernatant, and resuspend the pellet in 1 ml of PBS (or any buffer of choice, if for example the optimal lysis buffer has been already determined).
2. Sonicate with 3 mm probe at 20 % of intensity until the pellet is completely dissolved (do not overheat).
3. Take out 50 μ l, transfer into a fresh Eppendorf tube, and add SDS-PAGE loading dye (total extract).
4. Spin down the rest of the sample (950 μ l) for 3 min at maximum speed in table top centrifuge.
5. Take out 50 μ l, transfer into fresh Eppendorf tube, and add SDS-PAGE loading dye (soluble extract).
6. Use remaining 950 μ l to measure fluorescence (excitation: 488 nm, emission max: ~ 520 nm if a baculovirus expressing YFP such as EMBacY is used).
7. We strongly advice to measure the cellular fluorescence against a fluorescent standard in order to calibrate the measurements and to compare with expressions from other batches of virus.
8. Freeze the SNP/SN aliquots at -20 °C until use.

3.6.2 Using BIIC Stocks for Protein Production

(1 ml of BIIC for 1 l of expression culture.)

1. Quickly thaw one BIIC cryovial in a 37 °C water bath (or thaw in your hands, then use paper towels to protect your skin) with gentle agitation until cells are almost thawed.
2. Dilute the vial quickly into 100 ml insect cell medium.
3. Add this 100 ml suspension to 900 ml of uninfected Sf21 cells at a density of around 0.9×10^6 cells/ml.
4. Maintain cells at a density of 1×10^6 cells/ml until proliferation arrest.

5. Harvest cells at proper time after proliferation arrest (if EMBacY-based virus is used: when YFP reaches a plateau) and purify your protein.

4 Notes

1. Addition of penicillin (5–25 U/ml) and streptomycin (7 µg/ml) or gentamicin (5 µg/ml) can be useful when it is necessary to face a contamination.
2. Cell viability can be evaluated with Trypan blue. Mix one volume of cells with one volume of a 0.1 % stock solution of Trypan blue (in PBS or other isotonic salt solution). Nonviable cells will take up Trypan blue. Healthy cultures should contain more than 97 % of unstained viable cells. We use an automated counter that also provides cell size distribution, which is an indicator of cell infection.
3. In our experience, it is absolutely necessary to ascertain at this step that debris from the cell lysis are completely removed. Leftover debris, containing also genomic nucleic acid and RNA, can inhibit the transfection of insect cells by sequestering transfection reagent. PCR analysis can be performed with m13 Forward (CCCAGTCACGACGTTGTAAAACG) and m13 Reverse (CAGGAAACAGCTATGAC) primers. The size of amplified DNA fragment can be compared with the size of DNA fragment obtained from non-recombinant bacmid propagated by blue colonies. Column purified PCR reaction can be sequenced to check the integrity of inserted expression cassette.
4. We mainly use the bacmid BAC10:KO1629 [13] as a source of viral DNA. It consists of the wild-type AcMNPV genome in which part of ORF1629 that is essential for virus replication has been replaced by a low copy bacterial replicon and resistance makers (Kanamycin and Chloramphenicol). The bacterial replicon and resistance markers are surrounded by two Bsu36I restriction sites used for DNA linearization which increases recombination efficiency. ORF1629 is rescued after recombination with the transfer vector. Note that ready-to-use and genetically optimized linearized baculovirus DNA can be purchased from a number of sources (BaculoGold™ (BD Pharmingen), BacMagic™ and BacVector™ (Novagen), BacPAK6™ (Clontech), flashBAC™ (OET), Sapphire™ (Allele Biotechnology)).
5. Note that for initial transfection many other transfection reagents can be used as well. In our hands transfection with Fectofly (Polyplus), FuGENE, or X-Treme GENE HP (Roche) works with the same efficiency and offers the advantage that

the transfection suspension does not need to be removed after the 5 h incubation. Respect the incubation time recommended as extended incubation may lead to the formation of large and difficult to transfect DNA/transfection complexes. Carefully follow the manufacturer's instructions for procedures and the choice of compatible medium.

6. For amplification the multiplicity of infection (MOI) should be 0.1–0.4. Ideally the titer of the P0 stock should be determined experimentally. However, assuming a cell concentration of 0.5×10^6 cells/ml and a titer of the P0 stock of 2×10^7 pfu/ml, then an MOI of 0.1–0.4 would correspond to 0.25–1 % (virus volume/culture volume percentage). For example, add between 70 and 240 μ l of virus stock to a 25 ml culture in 250 ml flask.
7. Defective interfering particles (DIPs) are viral particles that miss part or all of their genome and thus cannot sustain an infection by themselves. Instead, they depend on coinfection with a suitable helper virus that provides the gene functions absent from the DIPs. Accumulation of DIPs that arise during passaging is favored by amplification at high MOIs when cells can be coinfecting with an intact virus and a DIP, allowing their replication. At low MOIs, formation of DIPs is limited as each cell is infected by a single viral particle. If the virus is defective, it will not replicate and DIPs will not accumulate [14, 16].
8. For PCR analysis of recombinant viruses, treat 200 μ l of viral suspension with 20 μ l of proteinase K at 20 mg/ml in 10 mM Tris-HCl (pH 8.0), 1 mM EDTA, 0.5 % SDS for 10 min at 72 °C and extract DNA with purification kit suitable for purification of large DNAs. This should provide enough DNA for running 100 PCR reactions.
9. For *in silico* design of oligonucleotides suitable for SLIC, Gibson or InFusion cloning uses SODA (<http://slc.cgm.cnrs-gif.fr/>) or NebBuilder (<http://nebuilder.neb.com/>).
10. One should consider factors that might affect the stability of the putative complex (pH, ionic strength, nonprotein ligand). Preparation of the cell lysate is a critical step that often requires optimization to identify a suitable lysis buffer. Optimal conditions should maximize the solubility and stability of the complex while minimizing oxidation, unwanted proteolysis, and aggregation. If the complexes can be tested *in vitro*, screening should include the use of a functional assay to control/optimize the activity of recombinant proteins.
11. Blunt-end ligation protocol allows insertion of the LoxP sequence in two possible orientations that will lead to distinct Cre-mediated fusion plasmids. A priori, there is no reason to privilege one orientation rather than the other, nevertheless

one cannot exclude that one orientation could be more favorable for virus stability.

12. For expression of two cDNAs, we would place one gene under the control of the PH promoter and another under that of promoter p10. For three, we would use two PH promoters and one p10 or vice versa. For expression of four cDNAs, use two PH and two p10 promoters.
13. If Cre/loxp concatenation between three plasmids (e.g., A, B, and C) does not work or interpretation is difficult (large plasmids and complex restriction patterns), proceed sequentially, i.e., isolate AB or BC fusion first and recombine with the third plasmid in a second step. Addition of Cre recombinase to a mixture of two plasmids leads to equilibrium after approximately 30 min at 37 °C. At equilibrium, only 20 % of initial plasmids are recombined (AB vector). For in silico modeling of a Cre-lox, recombination reaction uses a software Cre-ACEMBLER-1.0.1 (http://www.embl.fr/multibac/multiexpression_technologies/cre-acembler/.)

Acknowledgments

We thank Alice Aubert, Petra Drnkova, Maxime Chaillet, Isabelle Kolb, Natalie Troffer-Charlier, and Jean-Marie Garnier for sharing their experience on baculovirus expression and molecular biology. This work was funded by the CNRS, the INSERM, the Université de Strasbourg (UdS), the Alsace Region, and the French Infrastructure for Integrated Structural Biology (FRISBI) ANR-10-INSB-05-01 Instruct, part of the European Strategy Forum on Research Infrastructures (ESFRI) and supported by national member subscriptions. It benefited from grants ANR-12-BSV8-0015-01 from the Agence Nationale de la Recherche, INCA-2008-041 from the Institut National du Cancer, the Association pour la Recherche sur le Cancer, the Fondation pour la Recherche Médicale (FRM) (ING20101221017), and La Ligue contre le Cancer (fellowship to LR). IB acknowledges support from the European Commission (EC) Framework Programme (FP) 7 project ComplexINC (279039).

References

1. Baneyx F (1999) Recombinant protein expression in *Escherichia coli*. *Curr Opin Biotechnol* 10:411–421
2. Brondyk WH (2009) Selecting an appropriate method for expressing a recombinant protein. *Methods Enzymol* 463:131–147
3. Assenberg R, Wan P, Geisse S, Mayr L (2013) Advances in recombinant protein expression for use in pharmaceutical research. *Curr Opin Struct Biol* 23:393–402
4. Khan KH (2013) Gene expression in mammalian cells and its applications. *Adv Pharm Bull* 3:257–263

5. Picanco-Castro V, Biaggio RT, Cova DT, Swiech K (2013) Production of recombinant therapeutic proteins in human cells: current achievements and future perspectives. *Protein Pept Lett* 20:1373–1381
6. Fitzgerald DJ, Berger P, Schaffitzel C et al (2006) Protein complex expression by using multigene baculoviral vectors. *Nat Methods* 3:1021–1032
7. Berger I, Fitzgerald DJ, Richmond TJ (2004) Baculovirus expression system for heterologous multiprotein complexes. *Nat Biotechnol* 22:1583–1587
8. Bieniossek C, Imasaki T, Takagi Y, Berger I (2012) MultiBac: expanding the research toolbox for multiprotein complexes. *Trends Biochem Sci* 37:49–57
9. Sokolenko S, George S, Wagner A et al (2012) Co-expression vs. co-infection using baculovirus expression vectors in insect cell culture: benefits and drawbacks. *Biotechnol Adv* 30:766–781
10. Belyaev AS, Roy P (1993) Development of baculovirus triple and quadruple expression vectors: co-expression of three or four bluetongue virus proteins and the synthesis of bluetongue virus-like particles in insect cells. *Nucleic Acids Res* 21:1219–1223
11. Abdulrahman W, Uhring M, Kolb-Cheynelet I et al (2009) A set of baculovirus transfer vectors for screening of affinity tags and parallel expression strategies. *Anal Biochem* 385:383–385
12. Nie Y, Bellon-Echeverria I, Trowitzsch S et al (2014) Multiprotein complex production in insect cells by using polyproteins. *Methods Mol Biol* 1091:131–141
13. Zhao Y, Chapman DA, Jones IM (2003) Improving baculovirus recombination. *Nucleic Acids Res* 31:E6–6
14. Kool M, van Lier FLJ, Vlak JM, Tramper J (1991) Detection and analysis of *Autographa californica* nuclear polyhedrosis virus mutants with defective interfering properties. *Virology* 183:739–746
15. Vijayachandran LS, Viola C, Garzoni F et al (2011) Robots, pipelines, polyproteins: enabling multiprotein expression in prokaryotic and eukaryotic cells. *J Struct Biol* 175:198–208
16. Pijlman GP, van den Born E, Martens DE, Vlak JM (2001) *Autographa californica* baculoviruses with large genomic deletions are rapidly generated in infected insect cells. *Virology* 283:132–138

Chapter 6

Production of Cell Surface and Secreted Glycoproteins in Mammalian Cells

Elena Seiradake, Yuguang Zhao, Weixian Lu, A. Radu Aricescu, and E. Yvonne Jones

Abstract

Mammalian protein expression systems are becoming increasingly popular for the production of eukaryotic secreted and cell surface proteins. Here we describe methods to produce recombinant proteins in adherent or suspension human embryonic kidney cell cultures, using transient transfection or stable cell lines. The protocols are easy to scale up and cost-efficient, making them suitable for protein crystallization projects and other applications that require high protein yields.

Key words Protein expression, HEK 293 cells, Transient transfection, Stable cell lines

1 Introduction

About 30 % of the mammalian genome codes for secreted and membrane proteins, including more than half of all small-molecule drug-targets [1]. The majority of these proteins are recognized by the signal recognition particle (SRP) [2], which targets their translation to the endoplasmic reticulum (ER), from where they are trafficked to the Golgi apparatus and the cell membrane. The ER and Golgi contain specialized enzymes which mediate protein folding and posttranslational modification, such as glycosylation and lipidation [3–5]. The oxidizing environment in the ER also enables the formation of disulphide bridges between cysteine residues, a process that is essential for the correct folding of many eukaryotic extracellular proteins [6]. These proteins often fail to fold and mature correctly in prokaryotic cells, such as *Escherichia coli*, which lack ER and Golgi compartments. A widely used alternative to prokaryotic expression systems is baculovirus-driven

expression using insect cells [7]. Despite the many advantages of this system, the screening of multiple expression constructs is time- and labor-intensive, requiring the production of recombinant baculovirus for each construct. Another drawback is the lack of methods for controlling N-linked glycosylation in insect cells. Recent technological advances in using mammalian cells for the large-scale production of recombinant proteins have provided an attractive alternative offering rapid screening, methods to control N-linked glycosylation, and cost-efficient scaling up [8–12]. As a result, the structural characterization of secreted proteins (as measured by the increase of structures deposited in the Protein Data Bank) has increased exponentially [13]. The crystal structures for a first tranche of integral membrane proteins have now been solved using mammalian cell expression [14–17], heralding an increase also for this category of proteins.

Here we describe protocols developed for protein expression in Human Embryonic Kidney (HEK) 293 cells. They were originally developed for the production of secreted and membrane proteins [8, 10, 18], but can be applied to cytosolic and nuclear proteins. The specific cell lines used are HEK 293 cells expressing the SV40 large-T antigen (HEK293T), HEK 293S cells that lack *N*-acetylglucosaminyl transferase I activity and therefore do not produce complex N-linked glycans (HEK293S-GnTI⁻) [10, 19] and FreeStyle 293F (HEK293F), which are adapted to grow in suspension. We propose a versatile workflow (Fig. 1) that can be adapted to the expression of secreted or cell surface and membrane proteins. We suggest that the relevant protocol sections are studied before decisions are made on which materials are required for a specific project.

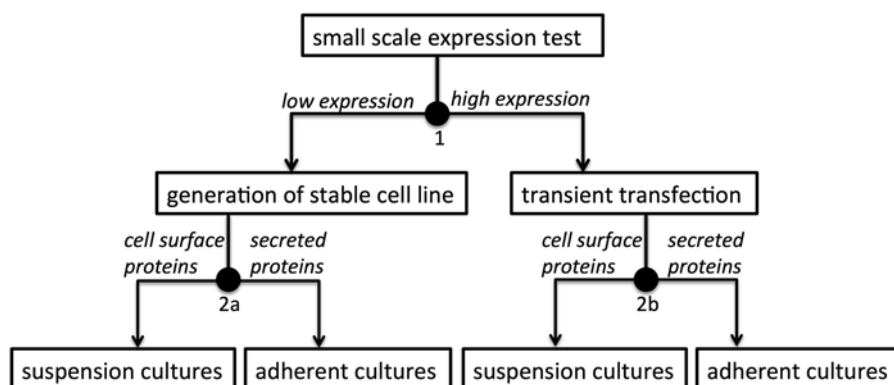


Fig. 1 Different strategies for protein expression are recommended, depending on the recombinant protein type and expression yields obtained. Decision points 1, 2a, and 2b are represented as *black circles*

2 Materials

2.1 Tissue Culture

1. Plastic tissue culture flasks (75 and 175 cm²).
2. Plastic roller bottles (expanded surface area 2,125 cm², Greiner 681070).
3. 96-, 6-, 12-well plates and 150×20 mm tissue culture dish.
4. Complete medium: DMEM with 10 % FBS, 1× glutamine, 1× nonessential amino acids. Store at 4 °C.
5. Low serum medium: DMEM with 2 % FBS, 1× glutamine, 1× nonessential amino acids. Store at 4 °C.
6. Serum-free medium: DMEM with 1× glutamine, 1× nonessential amino acids. Store at 4 °C.
7. FreeStyle 293 expression medium, 1× glutamine, 1× nonessential amino acids, 1 % FBS.
8. Optimem Glutamax medium (Lifetechnologies 11058021).
9. Hybridoma medium (Lifetechnologies 12045084).
10. Selenomethionine stock solution (30 mg/ml in serum-free DEMEM) filter sterilize and store in aliquots at −20 °C.
11. Methionine-free DMEM, 1× glutamine, 1× nonessential amino acids, 30 mg/l selenomethionine, 3 % dialysed FBS. Prepare immediately before using.
12. Kifunensine stock solution (1 mg/ml); dissolve solid in milliQ quality water warmed to 60 °C, filter to sterilize. Store at −20 °C.
13. Polyethyleneimine (PEI) stock solution (50 mg/ml) adjust pH to 7 with HCl (37 % v/v) and dilute to 1 mg/ml with water. Filter to sterilize. Store at −20 °C.
14. Phosphate buffer saline (PBS).
15. Trypsin-EDTA.
16. Polyethyleneimine (PEI) (Sigma 408727).
17. Polyethyleneimine max, (PEI-max) (Polyscience 24765).
18. X-tremeGENE HP DNA transfection reagent (Roche) or equivalent.
19. Kifunensine (Cayman chemical 10009437).
20. Selenomethionine (Eburon Organics 3211-76-5).
21. HEK293T cell line (ATCC CRL-3216).
22. HEK293S GnTI[−] cell line, *see* ref. 19.
23. HEK293F, Invitrogen FreeStyle 293F cell line (Life Technologies R790-07).

2.2 Other Reagents

1. Mammalian expression plasmids coding for the protein of interest (e.g. Fig. 2).
2. Integrase expression vector (pgk- ϕ C31/pCB92), *see* ref. 20.

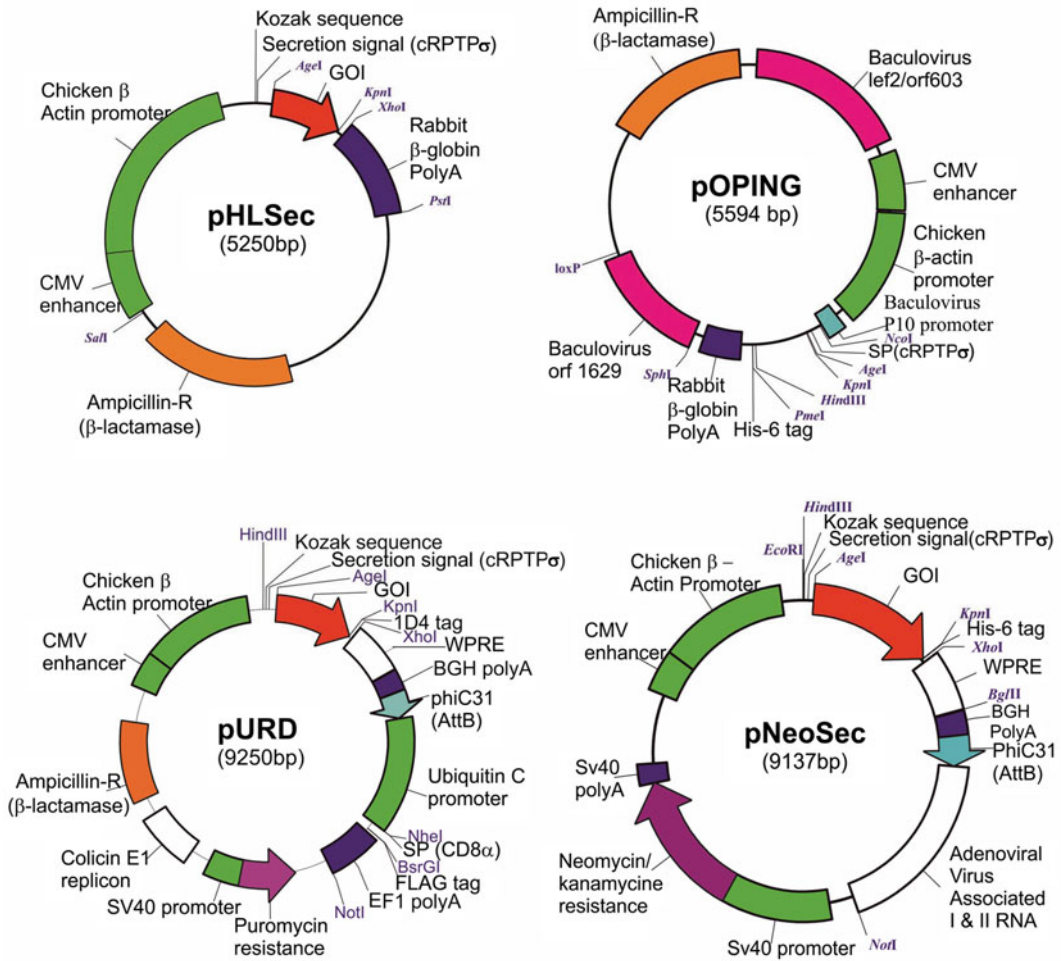


Fig. 2 Overview of popular mammalian expression vectors compatible with the protocols described. pHLSec [8] contains the cytomegalovirus (CMV) enhancer, a chicken beta actin promoter, a Kozak sequence, an optimized chicken protein receptor tyrosine phosphatase (RTP) sigma secretion signal, cloning sites, and a rabbit beta-globin polyadenylation (PolyA) sequence. pOPING [21] contains the same mammalian expression promoter, a secretion signal, and a rabbit beta-globin polyA. In addition, it contains a bacterial T7 promoter for bacterial expression and a baculovirus P10 promoter for insect/baculovirus gene expression. The vector has a standard baculovirus polyhedrin locus sequence. pURD [22] contains a mammalian cell selection cassette (SV40 promoter, puromycin resistance gene, PolyA). The primary expression cassette is essentially the same as in pNeoSec (see below), with a Rhodopsin 1D4 tag [28] at the C-terminus. The secondary expression cassette contains a human Ubiquitin C promoter, human CD8 alpha secretion signal, a FLAG tag [29] and a polyA sequence. pNeoSec [23] contains a phiC31 attachment site (AttB) for site-specific recombination with pseudo AttP sites in the mammalian cell genome when the PhiC31 integrase gene (pgk-phiC31/pCB92) is cotransfected, and a mammalian cell selection cassette (SV40 promoter, neomycin/G418 resistance gene, PolyA). The primary expression cassette contains the same promoter as pHLSec, a Woodchuck Hepatitis Virus posttranscriptional regulatory element (WPRE), and a bovine growth hormone (BGH) PolyA. An Adenovirus Virus-Associated RNA I and II sequence is included for binding the double-stranded RNA-activated protein kinase, DAI, thus suppressing the double stranded RNA response. The vector conveys kanamycin resistance for bacterial selection

3. Antibodies specific to the target protein or purification tag used (e.g. anti-pentaHis antibody, Qiagen 34660).
4. Standard SDS-page and western-blotting reagents.
5. Plasmid Giga kit (Qiagen 12191).

3 Methods

Carry out all procedures described in Subheadings 3.2–3.7 at room temperature, using ultra-pure water, sterile reagents, and under aseptic conditions unless otherwise specified.

3.1 Considerations When Designing Expression Constructs

There are a number of plasmid vectors available that are optimized for high levels of expression in mammalian cells. In this section, we describe protocols for transient expression using the vector suites pOPIN [21] and pHLSec [8], and for the generation of stable cell lines using the vectors pNeoSec or pURD [22, 23] (Fig. 2). The pOPIN vectors are also compatible with protein expression in bacterial and insect cell systems, useful for projects that require switching between different expression systems.

pOPING, pHLsec, and the derivative pNeoSec are high copy number plasmids and suitable for large-scale transient transfection requiring milligram quantities of plasmid DNA. pURD is a low-copy-number plasmid that enables coexpression of two inserted genes. We recommend using pURD for generating stable cell lines only. pOPING, pHLsec, pNeoSec, and pURD all contain a resident N-terminal signal sequence, which will direct the expressed protein to the secretory pathway. Care should be taken not to include both a native and the vector signal sequence. To check whether the target protein contains a signal sequence, we recommend using sequence analysis tools, such as the SignalP server [24] (<http://www.cbs.dtu.dk/services/SignalP>). Construct boundaries should be chosen to coincide with either the amino and carboxy-termini of the full length protein or subdomains to avoid miss-folding of incomplete domain fragments. It is often possible to predict structured subdomains through sequence alignment with proteins of known structure. Search engines that produce such sequence alignments are Fugue [25] (<http://tardis.nibio.go.jp/fugue>) and HHpred [26] (<http://toolkit.tuebingen.mpg.de/hhpred>) (see also Chapter 4).

3.2 Small-Scale Expression Test

Protein expression is first tested by transient transfection at small-scale (1–2 ml culture volume). Depending on the level of expression obtained, production is carried out by large-scale transient expression or via the construction of stable cell lines (Fig. 1, decision point 1).

1. *Day 1*: use a 175 cm² flask in which HEK293T cells are ~90 % confluent (Fig. 3a).

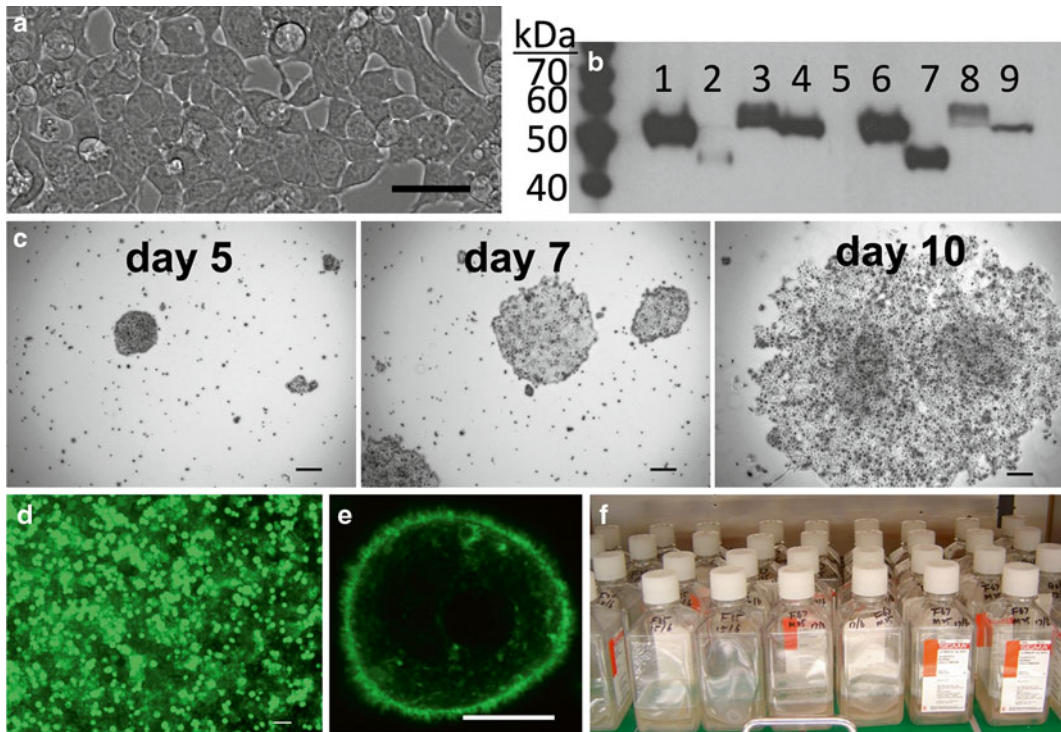


Fig. 3 (a) Adherent HEK 293T cells which are ~90 % confluent. Scale bar = 50 μm . (b) Analysis of small-scale expression tests for nine different secreted poly-histidine-tagged proteins. A protein ladder was loaded in the first lane, followed by nine different secreted protein test samples (8 μl medium supernatant, mixed with sample buffer, per lane). Sample 1 contains approximately 0.8 μg tagged protein. The proteins in lanes 1, 3, 4, 6–9 show yields >0.05 ng/ μl and will be expressed using transient transfection. We recommend expression in stable cell lines for protein 2 as it is weakly secreted. Protein 5 is not secreted. (c) Cells shown at different stages during stable cell line generation. 4–5 days after puromycin selection, the majority of cells have died and individual cell colonies start growing. The colonies are large enough to be cultured separately after 10 days. Scale bars = 50 μm . (d) Wide-field microscopy image showing suspension culture cells expressing an mVenus-tagged membrane protein. Scale bar = 50 μm . (e) Confocal microscopy image showing one cell expressing an mVenus-tagged membrane protein. Scale bar = 10 μm . (f) A suspension culture, grown in recycled square 500 ml medium bottles

2. Gently rinse the cells with 10 ml phosphate buffer saline (PBS) without detaching them. Remove the PBS and add 3 ml Trypsin-EDTA. Rock the flask gently to ensure all cells are covered by Trypsin-EDTA and incubate the flask at 37 $^{\circ}\text{C}$ for 5 min. Tap the sides of the flasks to help the cells detach from the plastic surface, add 7 ml of complete medium and resuspend the cells by gently pipetting up and down ten times with a 10-ml pipette.
3. Count the resulting cell suspension and adjust the concentration to 5×10^6 cells/ml by adding complete medium. Pipette 2 ml of the cell suspension into each well of a 6-well dish and incubate overnight at 37 $^{\circ}\text{C}$, 5 % CO_2 .

4. *Day 2*: the cells in the 6-well dish should be ~90 % confluent. Mix 4 µg of plasmid DNA and 100 µl Optimem Glutamax (or serum-free medium) by vortexing the tube.
5. Add 12 µl X-tremeGENE HP DNA transfection reagent (for a more economical but less efficient alternative, 1 mg/ml PEI may be used), vortex, and wait for 10 min. After 10 min, add 100 µl of the DNA mixture to the medium of a well in the 6-well dish and mix by swirling the plate gently.
6. Repeat the procedure for all constructs to be tested, using different wells in the 6-well dish. Incubate the plate for 2 days at 37 °C, 5 % CO₂.
7. *Day 4*: for expression of intracellular or membrane protein, remove cell media, and discard. Wash × 1 with PBS. Using a P-1000 µl pipette resuspend the adherent cells in 0.5 ml SDS gel loading buffer and sonicate the sample to break genomic DNA and reduce sample viscosity.
8. For secreted proteins, harvest the conditioned media and centrifuge for 10 min at 20,000 × *g* to pellet cell debris and then mix 50 µl of the resultant supernatant with 50 µl SDS gel loading buffer.
9. Analyze the samples by western blotting using antibodies specific to either the protein of interest or the purification tag contributed by the vector. If the expression is “low” (<0.05 ng/µl) or the amount of protein required exceeds 50 mg, move to Subheading 3.3. If the expression is high (>0.05 ng/µl) or only small amounts of protein are needed, for secreted proteins move to Subheading 3.4, for cell-surface and membrane proteins move to Subheading 3.5.

3.3 Generation of Stable Cell Lines

We describe the generation of stable cell lines (pools or clones) using either pNeoSec or pURD. Both vectors code for a selection marker and the phiC31 attachment site AttB to facilitate transgene integration into the host cell genome (Fig. 2). When using the pNeoSec vector, do not use HEK293T cells, as these are already resistant to neomycin (G418). Use HEK293S GnTI⁻ instead. If this has not been done already, we recommend testing vectors by small-scale transient expression prior to generating the stable cell lines.

1. *Day 1*. Use ~90 % confluent HEK293T or HEK293S GnTI⁻ cells, grown in 2 ml complete medium, in a well of a 6-well dish. Prepare the transfection cocktail by mixing 100 µl Optimem Glutamax (or serum-free medium) with 1 µg pNeoSec or pURD vector plasmid coding for the protein of interest, 3 µg integrase expression vector (pgk-φC31/pCB92) and 12 µl of X-tremeGENE HP DNA transfection reagent. Vortex briefly and let the mixture rest for 10 min.

2. Add the mixture to the cells. Mix by swirling the plate gently. Incubate overnight at 37 °C, 5 % CO₂.
3. *Day 2*. The cells should be ~100 % confluent. Gently rinse them with prewarmed PBS, add 300 µl Trypsin-EDTA, gently rock the plate to cover the cells with Trypsin-EDTA and incubate the plate at 37 °C for 5 min. Resuspend the cells by adding 1 ml of complete medium and gently pipetting up and down ten times. Transfer the cell suspension to a 75 cm² flask, add 15 ml complete medium. Incubate at 37 °C, 5 % CO₂.
4. *Day 5/6*. When the cells in the flask are 100 % confluent, add G418 (for pNeoSec) or puromycin (for pURD) up to final concentrations of 1 mg/ml or 2 µg/ml, respectively.
5. *Day 7–10*. A large fraction of the cells treated with puromycin die overnight. Remove the medium containing dead cells, add 10 ml fresh complete medium supplemented with 1 µg/ml puromycin. Incubate the flask for 3–10 days at 37 °C, 5 % CO₂. Cells treated with G418 die slower and need to be passaged: on day 7, rinse the cell carpet with prewarmed PBS, add 1 ml Trypsin-EDTA, incubate for 5 min at 37 °C, resuspend in 30 ml complete medium supplemented with 1 mg/ml G418, transfer the cell suspension to a 175 cm² flask and incubate the flask at 37 °C, 5 % CO₂. When the majority of cells have died (*see Note 1*), change the medium to fresh complete medium supplemented with G418 at 0.75 mg/ml. Incubate the flask at 37 °C, 5 % CO₂ for 5–10 days.
6. When individual cell clumps are growing on the flask surface, visible to the naked eye (Fig. 3c), rinse the cells with prewarmed PBS, add 3 ml Trypsin-EDTA, incubate for 5 min at 37 °C, resuspend the cells to in 12 ml (HEK293S GnTI⁻) or 25 ml (HEK293T) complete medium supplemented with 0.75 mg/ml (G418) or 1 µg/ml (puromycin). Transfer the resulting cell suspension to a 75 cm² (HEK293S GnTI⁻ cells) or 175 cm² (HEK293T cells) flask.
7. The cultures can be passaged to increase the number of cells using complete medium supplemented with 0.75 mg/ml G418 or 1 µg/ml puromycin and frozen using standard tissue culture protocols. After this initial selection period, the stable cell line can be grown without neomycin or puromycin.
8. For clonal selection, instead of transferring the cells to a 75 cm² flask on day 2, (**step 2** above), transfer the cells to a 150 × 20 mm tissue culture dish.
9. Add 25 ml complete medium and incubate at 37 °C, 5 % CO₂ for 3–4 days, or until the cells are ~90 % confluent. Add G418 (pNeoSec) or puromycin (pURD) at 1 mg/ml or 2 µg/ml, respectively.

10. Culture the cells for 7–14 days changing the medium (complete medium supplemented with 1 mg/ml G418 or 2 µg/ml puromycin) every 3–4 days. During this time, cell clumps (colonies), should become visible. Typical colony sizes are ~1–2 mm in diameter and visible by the naked eye (Fig. 3c).
11. Cut the tip of a 200-µl pipette and use this to transfer a single colony to a well in a 96-well tissue-culture plate. Do the same for all colonies, using a fresh tip for each. Culture the colonies with complete medium and test the expression levels by western blot analysis.
12. Expand cell lines as required for protein production. If the target is a secreted protein, follow the protocol in Subheading 3.4 for growing cells in attached culture. If the target is cell-surface-bound or a membrane protein, follow the procedure described in Subheading 3.6 for growing cells in suspension culture (Fig. 1, decision point 2a).

3.4 Cell Expansion in Adherent Culture

We describe a typical adherent cell expression experiment using 12 roller bottles. If necessary, adjust the number of bottles and reagents to match the protein yield required.

1. For each roller bottle, use a 175 cm² flask containing a 90–100 % confluent monolayer of HEK293T cells.
2. Detach the cells by rinsing them with 10 ml PBS, adding 3 ml Trypsin-EDTA, placing the flask at 37 °C for 5 min, tapping the flask gently, adding 7 ml complete medium and resuspending the cells by pipetting up and down ten times.
3. Transfer the cell suspension from one flask into one roller bottle and add 240 ml of prewarmed complete medium.
4. Place each bottle into a rotating incubator (for example, a Wheaton R2P Roller Apparatus) at 37 °C (*see Note 2*) for 3 days (HEK293T) or 5 days (HEK293S GnTI⁻). During this time, the cells will visibly attach to the plastic surface (*see Note 3*) and the medium will change to a lighter orange-red color.

If you are planning transient transfection (Fig. 1, decision point 2b), move to Subheading 3.5. If you are using a stable cell line, replace the medium in each bottle with 240 ml prewarmed low-FBS medium and place in a rotating incubator at 37 °C for 3–7 days. Supplement the low-serum medium with kifunensine (1 µg/ml final concentration) to manipulate N-glycosylation in HEK293T cells (*see Note 4*). Selenomethionine-labeled protein can be produced by replacing the culture medium with methionine-free medium containing 30 mg/ml selenomethionine (*see Note 5*).

3.5 Transient Transfection in Adherent Culture

1. For transfection of 12 roller bottles, mix 6 mg of plasmid DNA with 300 ml serum-free medium. Separately, mix 12 ml of PEI stock solution (1 mg/ml) with 300 ml of serum-free medium. Combine and mix the DNA and PEI solutions, and leave at room temperature for 10 min.
2. Carefully remove the medium from the roller bottles and discard it. Add 200 ml of prewarmed low-serum medium and 50 ml of the DNA/PEI mixture to each bottle and return bottles to the roller incubator.
3. The conditioned media are typically harvested after 3–5 days, when the desired level of protein expression has been achieved.

3.6 Cell Expansion in Suspension Culture

The following method describes the generation of suspension cultures from adherent HEK293S GnTI⁻ cells. Alternatively, use commercial HEK293F (*see* **Note 6**).

1. *Day 1*: use HEK293S GnTI⁻ cells grown to ~90 % confluent in a 175 cm² flask. Remove the medium and rinse the cells once with 10 ml FreeStyle 293 expression medium. Add 25 ml FreeStyle 293 expression medium and gently detach the cells by pipetting medium against the surface, using a 10 ml pipette (*see* **Note 7**).
2. Gently pipette the cell suspension up and down ten times and transfer the suspension to a 500 ml square bottle (*see* **Note 8**) containing 150 ml FreeStyle complete medium. Place the bottle on a sticky mat in an incubation shaker at 37 °C, 130 rpm, 8 % CO₂ for 3–4 days.
3. *Day 3–4*: dilute the cells to 2×10^5 cells/ml by adding FreeStyle complete medium and distribute 180-ml aliquots of the diluted cell suspension into an appropriate number of square 500-ml bottles. Place the bottle on a sticky mat in a shaking incubator at 37 °C, 130 rpm, 8 % CO₂ for 3–4 days. The cells can be expanded further after 3–4 days (they should reach cell densities of $2\text{--}3 \times 10^6$ /ml).

3.7 Transient Transfection in Suspension Culture

1. The following protocol is designed for ~1 L of suspension cells. Adjust volumes as required.
2. *Day 1*: follow the protocols described in Subheading 3.6 to prepare three square 500-ml bottles containing ~180 ml cell suspension each at $2\text{--}3 \times 10^6$ cells/ml.
3. Pool the cell suspensions and dilute by adding an equal volume (~540 ml) of FreeStyle complete medium. Distribute the diluted cell suspension into six square 500-ml bottles. Place these bottles into an incubation shaker at 37 °C, 130 rpm, 8 % CO₂.
4. *Day 2*: prepare the transfection cocktail for 1 liter of suspension cells: dilute 0.5 mg DNA in 12.5 ml of hybridoma medium.

Separately, dilute 1.5 ml of 1 mg/ml PEI-max in 12.5 ml of hybridoma medium. Combine the DNA and PEI-max solutions, mix and leave at room temperature for 5 min. Combine and centrifuge the suspension cells at $400\times g$ for 8 min (for example, in 550 ml Corning centrifuge tubes). Gently remove the medium without disturbing the cell pellet. Add the DNA/PEI-max solution to the cell pellet and resuspend the cells by pipetting gently up and down with a 25-ml pipette. Transfer and distribute the cell suspension into three square 500-ml bottles containing 170 ml prewarmed FreeStyle complete medium and place the bottles in the shaker at 37 °C, 130 rpm, 8 % CO₂ overnight.

5. *Day 3*: dilute the cell suspension in each bottle with an equal amount of FreeStyle complete medium and transfer half of the resulting suspension to a new bottle (maximum volume for each 500-ml bottle is 200 ml). The cells are typically harvested after 3–5 days, when the desired level of protein expression has been achieved (Fig. 3d, e).

4 Notes

1. If there is no sign of cell death after 2 days, use Trypsin-EDTA to resuspend and transfer half of the cells to each of two 175 cm² flasks. Add 25 ml fresh complete medium supplemented with 1 mg/ml G418 to each flask and incubate at 37 °C, 5 % CO₂ for several days.
2. There is no need for a 5 % CO₂ environment when growing cells in roller bottles.
3. The cells should cover at least 90 % of the plastic bottle surface after 3–4 days. The cell carpet may be uneven and include cell aggregates; this is no cause of alarm.
4. Simple N-linked glycans can be cleaved using recombinant Endo-F1, which is easily produced and purified from bacterial cells [27]. To inhibit the generation of complex N-linked glycans, which are not cleaved by Endo-F1, supplement the medium with 1 µg/ml kifunensine. Use kifunensine only when culturing the cells for protein expression, not while expanding the culture, as it reduces cell growth. Alternatively, use HEK293S GnTI⁻ cells which do not produce complex sugars, thus avoiding the need for kifunensine [10, 19].
5. To incorporate selenomethionine in recombinant protein expressed in adherent cultures, seed and grow cells as described in Subheadings 3.3–3.5. One day after transfection (transient expression) or 4 days after transferring cells into roller bottles (when using a stable cell line), prepare 250 ml of prewarmed

selenomethionine low-FBS medium per roller bottle. For each roller bottle, remove the medium and rinse the cell carpet by adding 50 ml of prewarmed PBS and rolling the bottle gently. Remove the PBS and repeat the rinsing step a second time. Add the selenomethionine low serum medium and place in a rotating incubator at 37 °C for 3–5 days, or until the cells start to detach.

6. On day 1, rapidly thaw an aliquot of HEK293F cells in a 37 °C water bath, transfer the cell suspension to a 15-ml Falcon tube containing 12 ml of FreeStyle complete medium and centrifuge at $400 \times g$ for 5 min. Remove the supernatant and resuspend the cells in 10 ml FreeStyle complete medium. Transfer the cell suspension to a 100 ml square bottle with 20 ml prewarmed FreeStyle complete medium and incubate at 37 °C, 130 rpm, 8 % CO₂ for 3–4 days. The resultant cell suspension can be expanded as described for HEK293S GnTI⁻ cells in Subheading 3.6, day 3–4.
7. It is important not to trypsinize at this stage, but to resuspend the cells mechanically.
8. The DMEM and other cell media used here are delivered in 500 ml square bottles. To reduce the costs, these bottles can be recycled to grow suspension cells after the medium is used up (Fig. 3f). For this purpose, ensure the inside of the bottle is kept sterile. Smaller bottles holding, for example, nonessential amino acids, can be used to culture smaller volumes.

References

1. Zheng C, Han L, Yap C et al (2006) Progress and problems in the exploration of therapeutic targets. *Drug Discov Today* 11:412–420
2. Grudnik P, Bange G, Sinning I (2009) Protein targeting by the signal recognition particle. *Biol Chem* 390:775–782
3. Komekado H, Yamamoto H, Chiba T, Kikuchi A (2007) Glycosylation and palmitoylation of Wnt-3a are coupled to produce an active form of Wnt-3a. *Genes Cells* 12:521–534
4. Braakman I, Bulleid NJ (2011) Protein folding and modification in the mammalian endoplasmic reticulum. *Annu Rev Biochem* 80:71–99
5. Aebi M (2013) N-linked protein glycosylation in the ER. *Biochim Biophys Acta* 1833:2430–2437
6. Ellgaard L (2004) Catalysis of disulphide bond formation in the endoplasmic reticulum. *Biochem Soc Trans* 32:663–667
7. Contreras-Gomez A, Sanchez-Miron A, Garcia-Camacho F et al (2014) Protein production using the baculovirus-insect cell expression system. *Biotechnol Prog* 30:1–18
8. Aricescu AR, Lu W, Jones EY (2006) A time- and cost-efficient system for high-level protein production in mammalian cells. *Acta Crystallogr D Biol Crystallogr* 62:1243–1250
9. Hacker DL, Kiseljak D, Rajendra Y et al (2013) Polyethyleneimine-based transient gene expression processes for suspension-adapted HEK-293E and CHO-DG44 cells. *Protein Expr Purif* 92:67–76
10. Chang VT, Crispin M, Aricescu AR et al (2007) Glycoprotein structural genomics: solving the glycosylation problem. *Structure* 15:267–273
11. Chaudhary S, Pak JE, Gruswitz F et al (2012) Overexpressing human membrane proteins in stably transfected and clonal human embryonic kidney 293S cells. *Nat Protoc* 7:453–466

12. Andrell J, Tate CG (2013) Overexpression of membrane proteins in mammalian cells for structural studies. *Mol Membr Biol* 30:52–63
13. Aricescu AR, Owens RJ (2013) Expression of recombinant glycoproteins in mammalian cells: towards an integrative approach to structural biology. *Curr Opin Struct Biol* 23: 345–356
14. Standfuss J, Edwards PC, D'Antona A et al (2011) The structural basis of agonist-induced activation in constitutively active rhodopsin. *Nature* 471:656–660
15. Standfuss J, Xie G, Edwards PC et al (2007) Crystal structure of a thermally stable rhodopsin mutant. *J Mol Biol* 372:1179–1188
16. Gruswitz F, Chaudhary S, Ho JD et al (2010) Function of human Rh based on structure of RhCG at 2.1 Å. *Proc Natl Acad Sci U S A* 107:9638–9643
17. Deupi X, Edwards P, Singhal A et al (2012) Stabilized G protein binding site in the structure of constitutively active metarhodopsin-II. *Proc Natl Acad Sci U S A* 109:119–124
18. Zhao Y, Bishop B, Clay JE et al (2011) Automation of large scale transient protein expression in mammalian cells. *J Struct Biol* 175:209–215
19. Reeves PJ, Callewaert N, Contreras R, Khorana HG (2002) Structure and function in rhodopsin: high-level expression of rhodopsin with restricted and homogeneous N-glycosylation by a tetracycline-inducible N-acetylglucosaminyltransferase I-negative HEK293S stable mammalian cell line. *Proc Natl Acad Sci U S A* 99:13419–13424
20. Chen CM, Krohn J, Bhattacharya S, Davies B (2011) A comparison of exogenous promoter activity at the ROSA26 locus using a PhiC31 integrase mediated cassette exchange approach in mouse ES cells. *PLoS One* 6:e23376
21. Berrow NS, Alderton D, Sainsbury S et al (2007) A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. *Nucleic Acids Res* 35:e45
22. Zhao Y, Malinauskas T, Harlos K, Jones EY (2014) Structural insights into the inhibition of Wnt signaling by cancer antigen 5T4/Wnt-activated inhibitory factor 1. *Structure* 22:612–620
23. Zhao Y, Ren J, Padilla-Parra S et al (2014) LIMP-2, the lysosome-sorting subunit of β -glucocerebrosidase, is targeted by the mannose 6-phosphate receptor. *Nat Commun* 5:4321
24. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340:783–795
25. Shi J, Blundell TL, Mizuguchi K (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol* 310:243–257
26. Soding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 33:W244–W248
27. Grueninger-Leitch F, D'Arcy A, D'Arcy B, Chene C (1996) Deglycosylation of proteins for crystallization using recombinant fusion protein glycosidases. *Protein Sci* 5:2617–2622
28. Wong JP, Reboul E, Molday RS, Kast J (2009) A carboxy-terminal affinity tag for the purification and mass spectrometric characterization of integral membrane proteins. *J Proteome Res* 8:2388–2396
29. Einhaue A, Jungbauer A (2001) The FLAG peptide, a versatile fusion tag for the purification of recombinant proteins. *J Biochem Biophys Methods* 49:455–465

Cell-Free Protein Synthesis Systems Derived from Cultured Mammalian Cells

Andreas K. Brödel, Doreen A. Wüstenhagen, and Stefan Kubick

Abstract

We present a technology for the production of target proteins using novel cell-free systems derived from cultured human K562 cells and Chinese hamster ovary (CHO) cells. The protocol includes the cultivation of cells, the preparation of translationally active lysates, and the cell-free synthesis of desired proteins. An efficient expression vector based on the internal ribosome entry site (IRES) from the intergenic region (IGR) of the cricket paralysis virus (CrPV) was constructed for both systems. The coupled batch-based platforms enable the synthesis of a broad range of target proteins such as cytosolic proteins, secreted proteins, membrane proteins embedded into endogenous microsomes, and glycoproteins. The glycosylation of erythropoietin demonstrates the successful performance of posttranslational modifications in the novel cell-free systems. Protein yields of approximately 20 µg/ml (K562-based cell-free system) and 50 µg/ml (CHO-based cell-free system) of active firefly luciferase are obtained in the coupled transcription-translation systems within 3 h. As a result, both cell-free protein synthesis systems serve as powerful tools for high-throughput proteomics.

Key words Cell-free protein synthesis, CHO cells, Coupled transcription-translation, Intergenic region internal ribosome entry site, In vitro protein production, K562 cells

1 Introduction

Cell-free protein synthesis systems based on eukaryotic extracts from wheat germ [1], yeast [2], rabbit reticulocytes [3], insect cells [4–7], and mammalian cells [8–10] have been developed in order to bypass the drawbacks of prokaryotic in vitro protein production. Until now, eukaryotic cell-free protein synthesis has been used to facilitate the production of a broad range of biotechnologically and pharmacologically relevant proteins, such as membrane proteins [6, 11], toxic proteins [12, 13], and antibody fragments [14, 15]. The functional characterization of target proteins in particular is in the scope of eukaryotic cell-free protein synthesis [16]. Each of the established eukaryotic cell-free systems offers advantages but also limitations in terms of yield, proper protein folding, posttranslational

modifications (PTMs), costs, scalability, speed, and ease of use [17]. Cell-free systems derived from cultured mammalian cells are of particular interest as mammalian cell lines are the preferred choice for the synthesis of biologically active proteins which require proper folding and PTMs [18]. However, the use of mammalian-based cell-free protein production systems has so far been limited due to significantly lower protein synthesis levels compared to prokaryotic *Escherichia coli*-based cell-free systems. Currently, the best performing eukaryotic in vitro platforms are derived from wheat germ [19] and *Spodoptera frugiperda* cells [4, 6, 20]. As a result, there is a high demand for the development of novel in vitro transcription-translation systems derived from cultured mammalian cell lines routinely used for industrial high-scale protein production. We have recently developed two novel cell-free systems derived from human K562 cells and CHO cells in order to address this demand [21–23]. Both coupled, batch-based platforms enable the production of a broad range of structurally and functionally divergent target proteins such as cytosolic proteins, secreted proteins, membrane proteins, and glycoproteins. Both cell-free systems support the PTMs, as has been demonstrated by the glycosylation of erythropoietin. Accordingly, these platforms contain translocationally active endogenous microsomes, enabling the incorporation of membrane proteins into biological membranes for functional studies. Both systems will further expand the use of cell-free protein synthesis as a suitable alternative to already established in vitro and in vivo platforms for structural and functional proteomics.

In this chapter, we present a detailed protocol for the synthesis of target proteins based on these cell-free systems derived from K562 and CHO cells. The protocol includes the cultivation of cells, the preparation of translationally active lysates, the construction of a suitable expression vector, and the cell-free synthesis of desired proteins.

2 Materials

2.1 Cell Culture

1. A conventional bioreactor (e.g., Biostat B-DCU II, Sartorius Stedim Biotech GmbH) or SuperSpinner (D 1000, Sartorius Stedim Biotech GmbH).
2. The chemically defined, serum-free Power CHO-2 CD medium (Lonza) supplemented with L-glutamine (Sigma-Aldrich).
3. ISF-1 serum-free medium (InVivo).
4. Wash and resuspension buffer: 40 mM HEPES-KOH (pH 7.5), 100 mM NaOAc, and 4 mM DTT. Store at 4 °C.

2.2 Preparation of Cell Lysates

1. Syringe with 20-gauge needle or high-pressure homogenizer.
2. Wash buffer (*see* Subheading 2.1, **item 4**).
3. Sephadex G-25 column (GE Healthcare).
4. S7 nuclease (Roche). Store at -20°C .
5. Nanodrop 2000c (Thermo Scientific).
6. 20 mM (20 \times) CaCl_2 .
7. 100 mM (15 \times) EGTA dissolved in water and stored at 4°C .

2.3 Generation of Expression Constructs

1. DNA template of the gene of interest (GOI).
2. DNA template of the CrPV IGR IRES (GenBank accession no. AF218039, nucleotides 6025–6216).
3. Expression vector EasyXpress pIX3.0 (Qiagen) or pcDNA3.1⁽⁺⁾ (Life Technologies) (*see* **Note 1**). Store at -20°C .
4. Primers for amplification of the GOI: *GOI-F* 5' ATG (N)₁₈ 3' and *GOI-R* 5' CTT GGT TAG GTT ATT (N)₂₁ 3' (*see* **Note 2**). Store at -20°C .
5. IRES-specific forward and reverse primer pair: *CrPV IGR IRES-F* 5' TTA AGA AGG AGA TAA ACA AAA GCA AAA ATG TGA TCT 3' (*see* **Note 3**) and *CrPV IGR IRES (ATG)-R* 5' (N)₁₅ CAT AGG TAA ATT TCT TAG GT 3' or *CrPV IGR IRES (GCT)-R* 5' (N)₁₅ AGC AGG TAA ATT TCT TAG GT 3' (*see* **Note 4**). Store at -20°C .
6. Adapter primers bearing the regulatory sequences (RS): *RS* 5' ATG ATA TCT CGA GCG GCC GCT AGC TAA TAC GAC TCA CTA TAG GGA GAC CAC AAC GGT TTC CCT CTA GAA ATA ATT TTG TTT AAC TTT AAG AAG GAG ATA AAC A 3' and *RS* 3' 5' ATG ATA TCA CCG GTG AAT TCG GAT CCA AAA AAC CCC TCA AGA CCC GTT TAG AGG CCC CAA GGG GTA CAG ATC TTG GTT AGT TAG TTA TTA 3'. Store at -20°C .
7. PCR cyclor and HotStar HiFidelity DNA Polymerase (Qiagen).
8. Restriction endonucleases (e.g., EcoRI and XhoI) and the corresponding restriction buffer solution (NEB). Store at -20°C .
9. QIAquick PCR Purification Kit (Qiagen).
10. PerfectBlue Gel system (Peqlab).
11. peqGOLD Universal-Agarose (Peqlab).
12. Rotiphorese 10 \times TBE buffer (Carl Roth).
13. Ethidium bromide (Carl Roth) or DNA Stain Clear G (SERVA).
14. 2-Log DNA Ladder (0.1–10.0 kb; NEB).
15. T4 DNA ligase and the corresponding reaction buffer solution (NEB). Store at -20°C .

16. Electrocompetent *E. coli* DH5 α cells. Store at -80°C .
17. Electroporation system and 1 mm cuvettes (Bio-Rad).
18. Super optimal broth with catabolite repression (SOC) medium. Store at -20°C .
19. Lennox broth (LB) medium (Carl Roth) and LB agar plates each with 100 $\mu\text{g}/\text{ml}$ ampicillin. Store at 4°C .
20. JETSTAR Plasmid Purification Kit (GENOMED).
21. Nanodrop 2000c (Thermo Scientific) used for the measurement of plasmid concentrations.
22. Primers for DNA sequencing: *M13-F* 5' GTA AAA CGA CGG CCA GTG 3', *M13-R* 5' CAG GAA ACA GCT ATG AC 3'. Primers anneal to the M13 region of pIX3.0-based plasmids. Store at -20°C .

2.4 Cell-Free Synthesis of Proteins

1. 25 mM (250 \times) amino acid mixture prepared by combining equal volumes of 25 mM stocks of each canonical amino acid (except tyrosine) dissolved in 50 mM HEPES containing 2 mM DTT and 25 mM tyrosine in 3 M KOH. Store the mixture at -20°C .
2. T7 RNA polymerase (Agilent). Store at -20°C .
3. 100 mM ATP, 100 mM CTP, 100 mM GTP, and 100 mM UTP stock solutions (Roche). Store at -20°C .
4. 500 mM (25 \times) creatine phosphate and 1 mg/ml (10 \times) creatine kinase is dissolved in water in each case and stored at -20°C .
5. 2.5 mM (10 \times) spermidine dissolved in water. Store at -20°C .
6. 25 mM (10 \times) DTT dissolved in water. Store at -20°C .
7. 1.5 M KOAc and 25 mM $\text{Mg}(\text{OAc})_2$ dissolved in water in both cases. Store at 4°C .
8. 3 mM (10 \times) m^7GpppG cap analogue dissolved in water. Store at -20°C .
9. Template DNA: pIX3.0-CrPV IGR IRES-GOI (*see* Subheading 3.3). Store at -20°C .
10. Thermomixer comfort (Eppendorf).

3 Methods

3.1 Cell Culture

1. Inoculate the fermenter or the SuperSpinner at a cell density of approximately 1.0×10^5 cells/ml.
2. CHO and K562 cells are grown at 37°C in batch mode and are harvested at a density of approximately 4.0×10^6 cells/ml (*see* **Note 5**).
3. Cells are collected by centrifugation at $200 \times g$, 4°C , for 5 min.

3.2 Preparation of Cell Lysates

1. Wash the cell pellet twice with the HEPES-based wash buffer.
2. Resuspend the cell pellet in the wash buffer to obtain a final cell density of approximately 1.0×10^8 cells/ml.
3. Cells are disrupted mechanically by passing the cell suspension through a 20-gauge needle using a syringe. Alternatively, cells can be disrupted by high-pressure homogenization.
4. Cell lysates are centrifuged at $10,000 \times g$, 4 °C, for 10 min in order to remove the nuclei and cell debris.
5. Supernatants are applied to a Sephadex G-25 column pre-equilibrated with wash buffer.
6. Collect the elution fractions (each 1 ml). Elution fractions harboring RNA with an absorbance above 100 at 260 nm are pooled.
7. Pooled elution fractions are treated with micrococcal nuclease in order to degrade residual mRNA. In this respect, 10 U/ml S7 nuclease and 1 mM CaCl_2 are added to the eluate and the reaction mixture is incubated for 2 min at room temperature. The reaction is inactivated by the addition of 6.7 mM EGTA (final concentration).
8. Cell lysates are immediately flash-frozen in liquid nitrogen and stored at -80 °C (*see Note 6*).

3.3 Generation of Expression Constructs

Procedures for the construction of efficient expression vectors follow conventional molecular cloning methods. The cloning strategy is illustrated in Fig. 1.

1. Amplify the coding sequences of the CrPV IGR IRES and the gene of interest in a two-step procedure using gene-specific and IRES-specific primers (*see Notes 7 and 8*).
2. Use the gene-specific reverse primer and the CrPV IGR IRES-specific forward primer in order to amplify and fuse the gene of interest downstream to the CrPV IGR IRES (*see Notes 7–9*).
3. Add regulatory sequences containing the cloning sites EcoRI and XhoI at the 5' and 3' noncoding regions of the final template using adapter primers in an additional PCR step (*see Notes 7 and 8*).
4. Digest the amplified PCR template and the EasyXpress pIX 3.0 vector with EcoRI and XhoI restriction nucleases using the recommended buffer and $1 \times$ BSA solution as described by the manufacturer NEB.
5. Purify the PCR-derived, restriction nuclease-digested template and the digested EasyXpress pIX 3.0 vector using the QIAquick PCR Purification Kit in order to remove disturbing small molecular mass components.

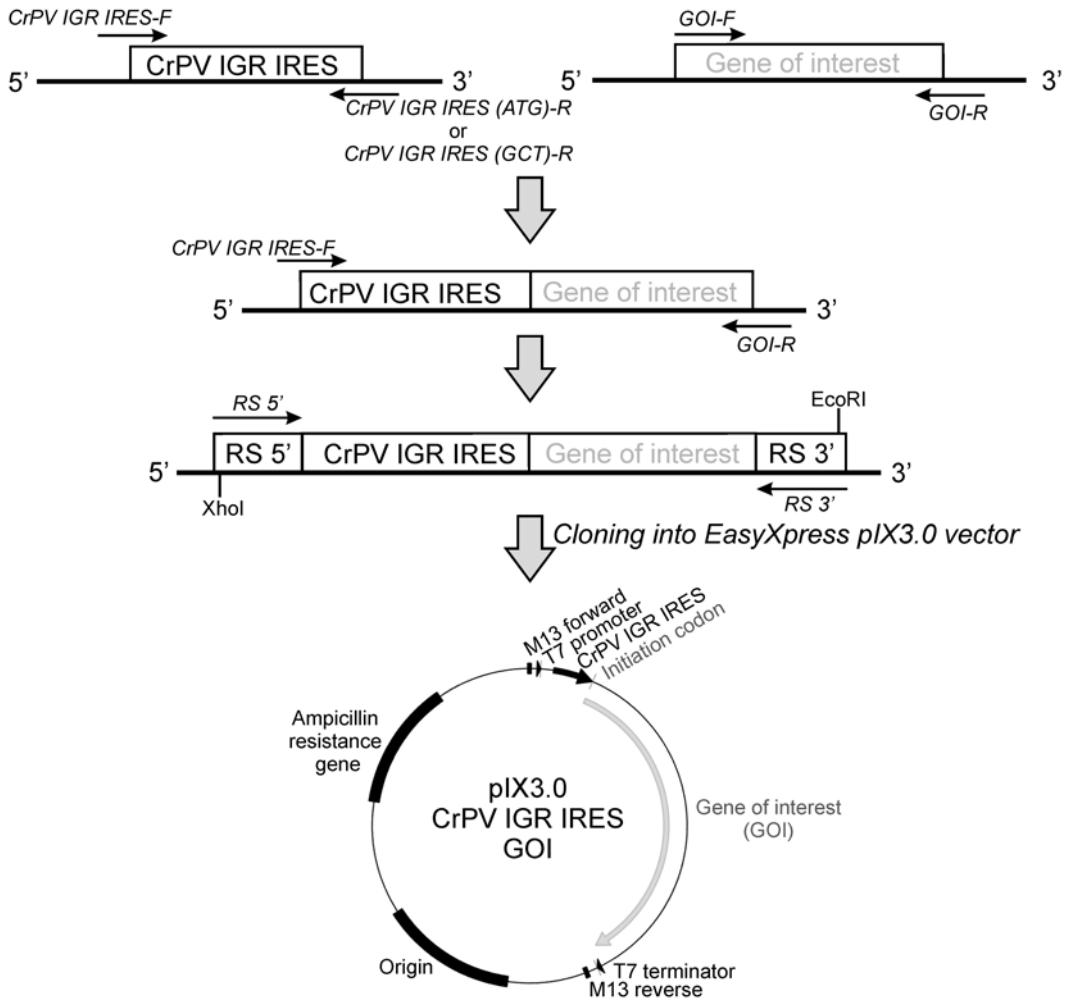


Fig. 1 Construction of the expression vector EasyXpress pIX3.0-CrPV IGR IRES-GOI. The CrPV IGR IRES is cloned upstream of the gene of interest (GOI) encoding sequence in the EasyXpress pIX3.0 vector backbone. Replacement of the AUG to a GCU initiation codon leads to increased protein synthesis levels

- Determine the concentration of the purified samples using the Nanodrop 2000c.
Use the purified DNA template and the purified, linearized expression vector in a molar ratio of 2:1 to 6:1 in order to ligate both DNA fragments. Ligation reactions are prepared according to the manufacturer's recommendations (NEB) and are incubated for 2 h at 20 °C in a thermomixer.
- Transform the ligation product into *E. coli* cells. For this purpose, 20 µL of the electrocompetent *E. coli* cell suspension (strain DH5α) is mixed with 1 µL of the ligation reaction.

Incubate the mixture on ice for 1 min and transfer the sample to precooled electroporation cuvettes. Electroporation is performed at 1.8 kV and the cell suspension is mixed with 1 ml SOC medium.

8. Incubate the mixture for 30–60 min at 37 °C and streak an aliquot of 200 µL onto an LB agar plate containing 100 µg/ml ampicillin.
9. Incubate the agar plate at 37 °C overnight.
10. Cultivate single colonies in LB medium containing 100 µg/ml ampicillin at 37 °C overnight.
11. Extract the plasmid using the JETSTAR Plasmid Purification Kit according to the manufacturer's recommendations (GENOMED).
12. Determine the concentration of the extracted plasmid DNA.
13. Confirm the nucleotide sequence of the cloned construct by DNA sequencing using the primers *M13-F* and *M13-R*.

3.4 Cell-Free Synthesis of Proteins

1. Thaw all necessary compounds including the cell lysate and store them on ice (*see Note 10*).
2. Coupled transcription-translation reactions in 25 µl batch volumes (*see Note 11*) are composed of 40 % lysate, 100 µM of each canonical amino acid, nucleoside triphosphates (1.75 mM ATP, 0.30 mM CTP, GTP and UTP, respectively), 20 mM creatine phosphate, 100 µg/ml creatine kinase, 20 nM vector DNA, 0.25 mM spermidine, 2.5 mM DTT, 1 U/µl T7 RNA polymerase (*see Notes 12 and 13*), and 0.3 mM m⁷GpppG (*see Note 14*). The final concentrations of KOAc and Mg(OAc)₂ depend on the source of the cell lysate. For CHO-based protein synthesis, 3.9 mM Mg(OAc)₂ and 150 mM KOAc are used. In contrast, 3.4 mM Mg(OAc)₂ and 120 mM KOAc are used in the K562-based cell-free system. Detection and analysis of the synthesized protein can be performed by the addition of radioisotope-labeled amino acids (e.g., ¹⁴C-labeled-leucine, specific radioactivity 46–75 dpm/pmol).
3. Protein synthesis is operated for 3 h in a thermomixer with shaking at 500 rpm and 33 °C (*see Note 15*).
4. Store samples immediately on ice after completion of the cell-free reaction. For long-term storage, freeze samples in liquid nitrogen and keep them at –20 °C until further analysis.
5. The performance of the eukaryotic cell-free systems was evaluated by synthesizing a broad range of target proteins such as the cytosolic proteins firefly luciferase (LUC) and enhanced yellow fluorescent protein (eYFP), the secreted protein

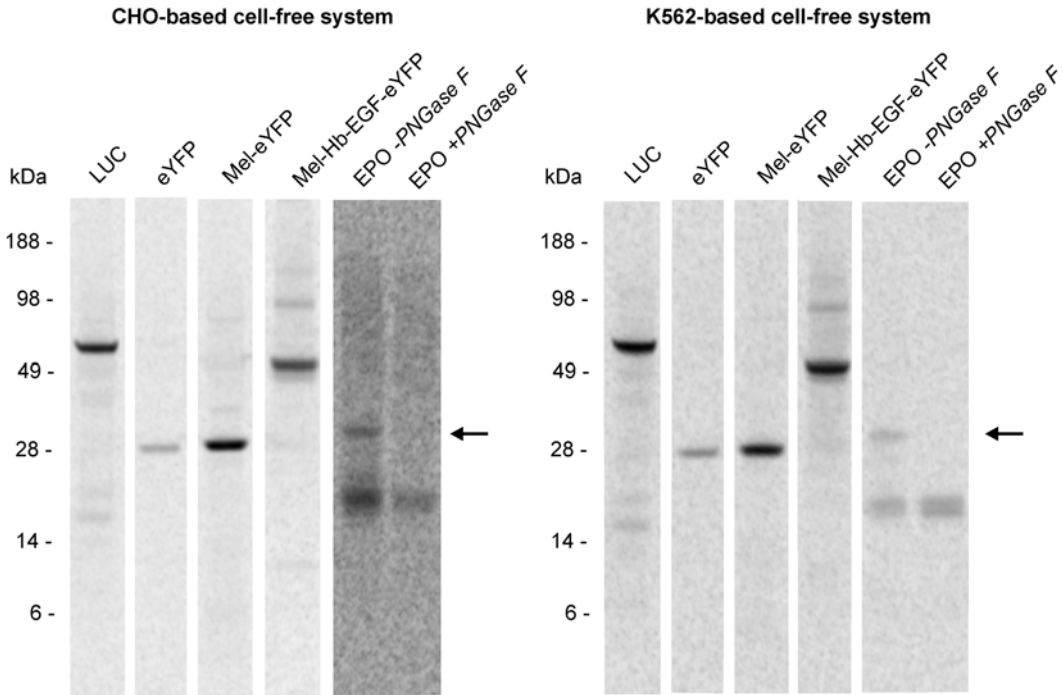


Fig. 2 Qualitative analysis of ^{14}C -leucine-labeled proteins synthesized in CHO- and K562-based cell-free systems. Cell-free synthesis of cytosolic proteins LUC (61 kDa) and eYFP (27 kDa), secreted protein Mel-eYFP (29 kDa), membrane protein Mel-Hb-EGF-eYFP (51 kDa), and glycoprotein EPO (21 kDa, unglycosylated). Integrity of target proteins was visualized by autoradiography after gel electrophoresis (Typhoon Trio+ Imager, GE Healthcare). EPO is presented before (–) and after (+) deglycosylation with PNGase F. *Arrows indicate glycosylated EPO*

Mel-eYFP, the type I transmembrane protein heparin-binding EGF-like growth factor (Mel-Hb-EGF-eYFP), and the glycoprotein erythropoietin (EPO) (Fig. 2). In the case of Mel-Hb-EGF-eYFP, the native signal peptide of Hb-EGF was replaced by the melittin signal sequence (Mel). Additionally, eYFP was fused to the C-terminus of this type I membrane protein in order to facilitate microscopic investigations. In the case of Mel-eYFP, the melittin signal sequence was added to the N-terminus of the eYFP in order to enforce protein translocation into microsomes present in the eukaryotic cell lysates. Typically, these microsomes have a diameter of 0.5–4.0 μm and tend to form large aggregates. Protein translocation into endogenous microsomes was investigated by CLSM analysis (LSM 510, Zeiss) of cell-free synthesized Mel-eYFP and Mel-Hb-EGF-eYFP (Fig. 3). CLSM images were analyzed using the Zeiss LSM Imaging software.

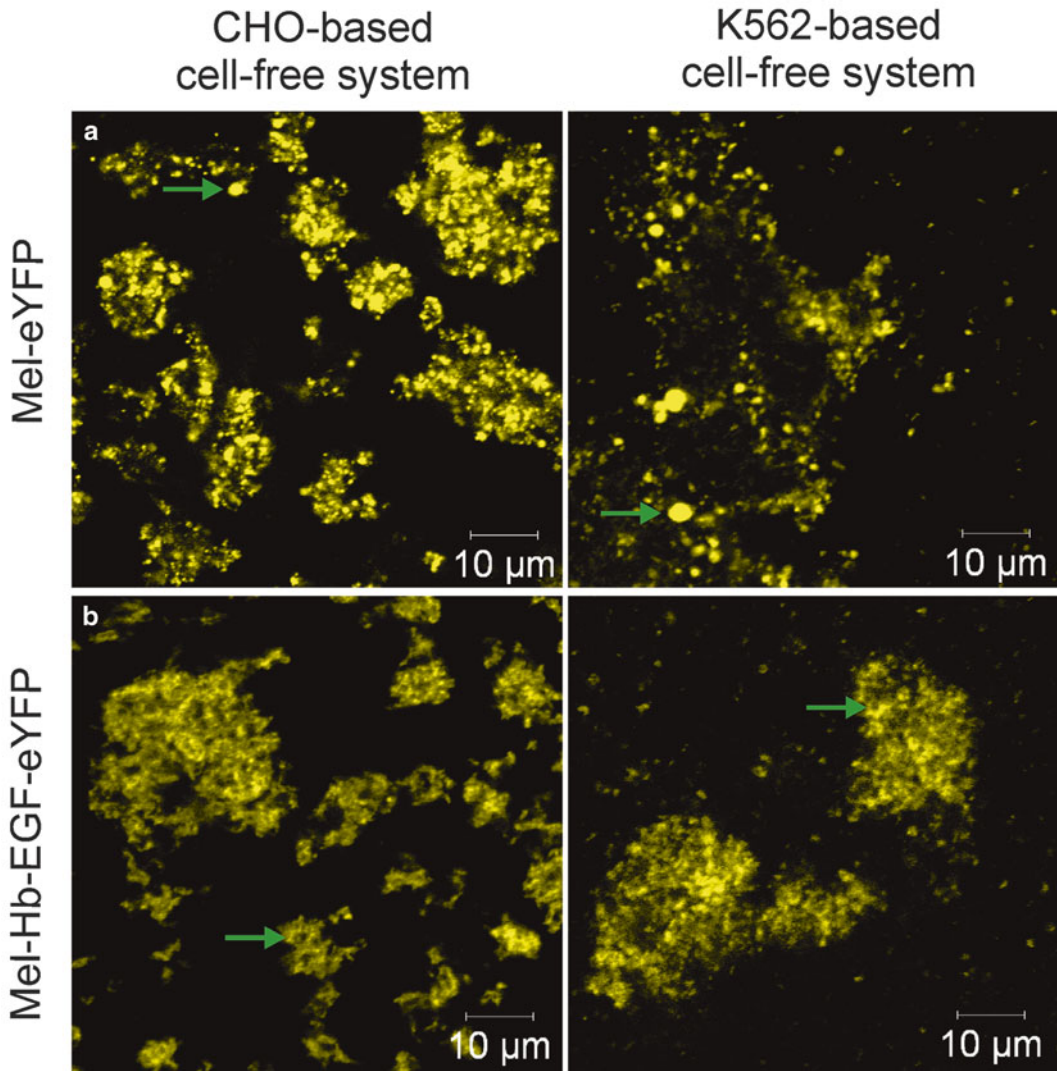


Fig. 3 CLSM analysis of the cell-free synthesized secreted protein Mel-eYFP and the type I membrane protein Mel-Hb-EGF-eYFP in CHO-based and K562-based expression systems. **(a)** Fluorescent vesicles indicate the synthesis and translocation of Mel-eYFP into the lumen of the microsomes. **(b)** Microsomes with fluorescent membranes visualized by CLSM indicate the production and insertion of the Mel-Hb-EGF-eYFP into microsomal membranes. eYFP was excited at 488 nm and emission was monitored with a long-pass filter in the wavelength range above 505 nm. Arrows exemplarily indicate individual microsomal vesicles harboring the fluorescent target protein. Microscope settings were individually adjusted for each sample as fluorescent intensities differed

4 Notes

1. EasyXpress pIX3.0-CrPV IGR IRES and pcDNA3.1⁽⁺⁾-CrPV IGR IRES constructs can both be used for cell-free protein production. pcDNA3.1-based constructs equipped with the

CrPV IGR IRES usually enable the synthesis of lower amounts of active LUC compared to the pIX3.0-CrPV IGR IRES vector. However, pcDNA3.1⁽⁺⁾-CrPV IGR IRES (ATG)-based constructs can be applied for the synthesis of target proteins in vitro and in vivo without recloning.

2. Reverse primers necessarily need an overlapping region (underlined sequence) at their 3' end in order to anneal to the adapter primer *RS* 3'. N indicates the base sequence of the gene of interest.
3. The *CrPV IGR IRES-F* primer necessarily needs an overlap (underlined sequence) at its 5' end in order to anneal to the adapter primer *RS* 5'.
4. Protein synthesis levels can be further increased by the replacement of the initiation codon AUG to GCU [21]. For this purpose, the *CrPV IGR IRES (GCT)-R* primer has to be used instead of the primer *CrPV IGR IRES (ATG)-R*. Both CrPV IGR IRES-R primers necessarily require an overlapping region (underlined sequence) to the respective sequence of the gene of interest.
5. Harvest cells at the late-log phase in order to obtain lysates with highest translational activity.
6. Aliquot cell lysates in order to avoid repeated freeze-thaw cycles.
7. Use the proofreading HotStar HiFidelity DNA Polymerase in order to minimize mutations.
8. Analyze 1 μ L of each PCR product on a 1 % (w/v) agarose gel stained with ethidium bromide or SERVA DNA Stain Clear G.
9. Increased protein synthesis levels are obtained utilizing the CrPV IGR IRES.
10. Assure that all compounds are completely thawed and fully dissolved before use.
11. Cell-free reactions are scalable according to the desired application. Reported volumes are in the range of several nanoliters to a 100 L [24–27].
12. Any nuclease contamination must be avoided. DNase/RNase-free water and filter pipet tips have to be used. Wear gloves as a barrier for contaminations.
13. Gently vortex all reaction compounds before use.
14. Addition of m⁷GpppG cap analogue is not essential for protein synthesis but leads to increased protein production levels.
15. Reaction temperature has been adapted to the synthesis of active firefly luciferase. It has to be noticed that optimal reaction temperatures for the production of other target proteins may vary slightly [28].

Acknowledgment

This research is supported by the German Ministry of Education and Research (BMBF Nos. 0312039 and 0315942).

References

1. Madin K, Sawasaki T, Ogasawara T et al (2000) A highly efficient and robust cell-free protein synthesis system prepared from wheat embryos: plants apparently contain a suicide system directed at ribosomes. *Proc Natl Acad Sci U S A* 97:559–564
2. Hodgman CE, Jewett MC (2013) Optimized extract preparation methods and reaction conditions for improved yeast cell-free protein synthesis. *Biotechnol Bioeng* 110:2643–2654
3. Martoglio B, Dobberstein B (2006) Cotranslational translocation of proteins into microsomes derived from the rough endoplasmic reticulum of mammalian cells. In: Celis JE (ed) *Cell biology: a laboratory handbook*. Academic, New York, pp 215–221
4. Kubick S, Schacherl J, Fleischer-Notter H et al (2003) *In vitro* translation in an insect-based cell-free system. In: Swartz J (ed) *Cell-free protein expression*. Springer, Berlin, pp 209–217
5. Kubick S, Gerrits M, Merk H et al (2009) *In vitro* synthesis of posttranslationally modified membrane proteins. In: Larry D (ed) *Current topics in membranes*, vol 63(2). Academic, New York, pp 25–49
6. Sachse R, Wüstenhagen D, Šamálíková M et al (2012) Synthesis of membrane proteins in eukaryotic cell-free systems. *Eng Life Sci* 13:39–48
7. Royall E, Woolaway KE, Schacherl J et al (2004) The *Rhopalosiphum padi* virus 59 internal ribosome entry site is functional in *Spodoptera frugiperda* 21 cells and in their cell-free lysates: Implications for the baculovirus expression system. *J Gen Virol* 85:1565–1569
8. Mikami S, Kobayashi T, Imataka H (2010) Cell-free protein synthesis systems with extracts from cultured human cells. *Methods Mol Biol* 607:43–52
9. Mikami S, Masutani M, Sonenberg N et al (2006) An efficient mammalian cell-free translation system supplemented with translation factors. *Protein Expr Purif* 46:348–357
10. Zeenko VV, Wang C, Majumder M et al (2008) An efficient *in vitro* translation system from mammalian cells lacking the translational inhibition caused by eIF2 phosphorylation. *RNA* 14:593–602
11. Nozawa A, Ogasawara T, Matsunaga S et al (2011) Production and partial purification of membrane proteins using a liposome-supplemented wheat cell-free translation system. *BMC Biotechnol* 11:35
12. Orth JHC, Br S, Boundy S et al (2010) Cell-free synthesis and characterization of a novel cytotoxic pierisin-like protein from the cabbage butterfly *Pieris rapae*. *Toxicon* 57:199–207
13. Bechlars S, Wüstenhagen DA, Dräger K et al (2013) Cell-free synthesis of functional thermostable direct hemolysins of *Vibrio parahaemolyticus*. *Toxicon* 76:132–142
14. Stech M, Merk H, Schenk JA et al (2012) Production of functional antibody fragments in a vesicle-based eukaryotic cell-free translation system. *J Biotechnol* 164:220–231
15. Merk H, Gless C, Maertens B et al (2012) Cell-free synthesis of functional and endotoxin-free antibody Fab fragments by translocation into microsomes. *Biotechniques* 53:153–160
16. Brödel AK, Raymond JA, Duman JG et al (2013) Functional evaluation of candidate ice structuring proteins using cell-free expression systems. *J Biotechnol* 163:301–310
17. Braun P, LaBaer J (2003) High-throughput protein production for functional proteomics. *Trends Biotechnol* 21:383–388
18. Jayapal K, Wlaschin K, Hu W et al (2007) Recombinant protein therapeutics from CHO cells - 20 years and counting. *Chem Eng Prog* 103:40–47
19. Endo Y, Sawasaki T (2006) Cell-free expression systems for eukaryotic protein production. *Curr Opin Biotechnol* 17:373–380
20. Ezure T, Suzuki T, Higashide S et al (2006) Cell-free protein synthesis system prepared from insect cells by freeze-thawing. *Biotechnol Prog* 22:1570–1577
21. Brödel AK, Sonnabend A, Kubick S (2014) Cell-free protein expression based on extracts from CHO cells. *Biotechnol Bioeng* 111:25–36
22. Stech M, Brödel AK, Sachse R et al (2013) Cell-free systems: functional modules for synthetic and chemical biology. In: Scheper T, Belkin S, Doran PM et al (eds) *Advances in biochemical engineering biotechnology*, vol 137. Springer, Berlin, pp 67–102

23. Brödel AK, Sonnabend A, Roberts LO et al (2013) IRES-mediated translation of glycoproteins and membrane proteins in eukaryotic cell-free systems. *PLoS One* 8:e82234
24. Angenendt P, Nyarsik L, Szaflarski W et al (2004) Cell-free protein expression and functional assay in nanowell chip format. *Anal Chem* 76:1844–1849
25. Voloshin AM, Swartz JR (2008) Large-scale batch reactions for cell-free protein synthesis. In: Spirin AS, Swartz JR (eds) *Cell-free protein synthesis*. Wiley, Weinheim, pp 207–235
26. Swartz J (2006) Developing cell-free biology for industrial applications. *J Ind Microbiol Biotechnol* 33:476–485
27. Stoevesandt O, Taussig MJ, He M (2009) Protein microarrays: high-throughput tools for proteomics. *Expert Rev Proteomics* 6: 145–157
28. Spirin AS, Swartz JR (2008) Cell-free protein synthesis systems: historical landmarks, classification, and general methods. In: Spirin AS, Swartz JR (eds) *Cell-free protein synthesis*. Wiley, Weinheim, pp 1–34

Crystallization: Digging into the Past to Learn Lessons for the Future

Vincent J. Fazio, Thomas S. Peat, and Janet Newman

Abstract

Crystals of biological macromolecules have been observed and grown for well over a century. More effort has been put into biological crystallization in the last few decades due to the importance of X-ray crystal structures, the advent of synchrotron radiation sources, improved computational speed, better software, and the availability of recombinant protein. Here we focus on two important areas of crystal growth: firstly, on techniques for stabilizing the protein sample, and secondly, on strategies and approaches for selecting the crystallization cocktails most suitable for different strategies.

Key words Crystallization screening, Proteins, Differential scanning fluorimetry

1 Introduction

Crystals of macromolecules have been deliberately grown for well over 100 years; in particular, crystals of different hemoglobins have been grown since the mid-nineteenth century [1]. Initially, proteins were crystallized simply because they could be, but it was certainly recognized that crystallization was an efficient purification mechanism [2]. Generally, proteins are crystallized now in order to elucidate their three-dimensional structure, and countless millions of crystallization experiments have been set up to this end. Despite the vast numbers of crystallization trials that have been performed and the impressive number of structures available from X-ray crystallography (over 85,000 from the Protein Data Bank (PDB) [3]; <http://www.rcsb.org>), it does not appear on first glance that we have achieved the improvements in crystallization efficiency over time that one might expect. Consider gene sequencing—which might be viewed as an analogous enabling technology. The first gene sequence was published in 1972 [4], and in 1977, the first full DNA genome of the bacteriophage ϕ X174 (which contains 5×10^3 base pairs) was published by the Sanger group [5]. Within 30 years, the complete sequence of a human (3×10^9 bp)

was available [6, 7] and by the early 2000s, the move away from Sanger dideoxy chain termination sequencing [8] to “next-generation” technologies [9–11] had a profound effect on the efficiency of sequencing (*see*, e.g., <http://www.genome.gov/sequencingcosts/>). The state of sequencing is now that it costs just cents per kilobase; given sufficient sample (conservatively, nanograms), a sequence can be obtained by any number of standard methods. We are still a long way away from that in crystallization—we estimate each crystallization trial costs between \$0.01 and \$1.00 to set up—excluding the cost of preparing the sample, and generally hundreds to thousands of experiments are set up for any given protein. At least several micrograms of pure sample are required, and we cannot guarantee a crystal—of diffraction quality or not—at the end of the process. Like sequencing, there are many technologies that can be used, for example, vapor diffusion or microbatch, but in the case of crystallization, the different techniques won’t necessarily give the same result.

The lack of parallels to gene sequencing does not mean that no progress has been made; there have been a number of breakthroughs in macromolecular crystallization and the utilization of those crystals for X-ray diffraction analysis, including a Nobel Prize in 1946 for the preparation of pure enzymes and virus proteins. Other highlights include showing that protein crystals can diffract X-rays [12], the development of the vapor diffusion method [13], the use of polyethylene glycol as a precipitating agent [14, 15], the development of molecular biology to allow for the large-scale heterologous expression of proteins [16], the development of affinity tags for purification [17–20], the development of the first screening kit [21], the development of cryo-crystallography [22, 23], and the development of crystallization automation [24–27]. Additionally, micro-focus beamlines [28] have been developed to deal with ever-smaller crystals. However, despite over 150 years’ experience and despite all of the breakthroughs, we still don’t know how to form crystals of any given protein, or indeed, if it will crystallize at all.

Crystallization of a protein is an interplay between many different factors including the molecule itself, how the molecule is formulated, the availability of the sample, the crystallization technique, the crystallization conditions, and the physical parameters of the experiment [13]. At least one of these aspects will be limiting, most often the amount of the protein sample, but the bottleneck might be the experimentalist’s access to different temperatures for incubation, or different chemicals, the solubility of the protein, the availability of automation, and so on. More knowledge would make the inevitable choices easier—if we knew which areas of chemical space were intrinsically more likely to produce crystals, we would choose initial screens that probe those hotspots. Given that there is a real and significant cost to each crystallization trial

that is set up, we also need a good metric to tell us when to stop and when to start. There are two questions that need to be addressed: First, does this sample have any chance of crystallizing at all? Ideally, this would be a reliable computational test requiring only information about the primary sequence of the construct under consideration. There are computational tools that address this question (e.g., XtalPred [29, 30]), but to date these have not been enthusiastically embraced by the experimentalists. The second question comes up when some crystallization trials have already been set up—what can we glean from this information to get some statistical measure of the likelihood of success given the lack of it so far? In the absence of such tests, what analyses can be done to answer a project's stop/go question? An unfortunate result of the variation in the limiting factors in different crystallization laboratories is that there are no general rules or strategies for the optimal approach to setting up crystallization trials, however, there are some factors which can (or should) be used in any crystallization campaign. In this paper we will discuss different approaches to crystallization screening, given different limitations. Most of the techniques used in crystallization today have been around for decades; the state of the art is not in new techniques per se, but in the judicious application of the existent tools.

2 The Protein Sample

The protein sample is undoubtedly the most important variable in a crystallization experiment [31]; it is the constant between the individual trials that make up a crystallization campaign. Working with the sample gives the experimentalist an initial feeling for the stop/go decision. Getting the right sample depends not only on having the right construct (sequence, extent, tags) [32] but also ensuring that the protein product from the construct is well behaved. What does that entail? The protein sample should be pure (free from other contaminating molecules including different states of posttranslational modification), conformationally homogeneous (all molecules in the same pose), and in as minimal formulation as possible. To fulfill this requires that the protein be well folded: the requirement for conformational homogeneity demands this. Aggregation, precipitation, and the requirement for high concentration of salt or cosmotrophs (e.g., glycerol) are indications that the protein sample does not fulfill these basic requirements. Dynamic light scattering (DLS) is a noninvasive biophysical technique that measures the average hydrodynamic radius of particles in solution [33]. An aggregate will dominate a DLS measurement and is an indication that some modification to the protein sample might be appropriate. The treatment could be as minor as filtering the sample; subsequent DLS measurement will show if more interventions

are needed. Despite DLS being quite a readily available technique which uses only a small amount of (recoverable) sample and is quick to run, it is not particularly widely used as a general pre-crystallization checkpoint. There are literature references to the utility of DLS in macromolecular crystallization dating back at least 15 years [34]; however, in our experience, DLS is generally only used after initial crystallization trials have proven recalcitrant. Although previous work has suggested that monodisperse protein samples are more likely to produce crystals than samples containing aggregate [35], once a protein sample has been created, it is often just as easy to set up the sample in crystallization trials as it is to modify the sample in order to get a better DLS signal. This introduces one of the truisms of protein crystallization: no high-throughput (or low-volume) solution-based analytical technique is yet known that has a strong positive correlation with crystallizability of the sample. As our primary goal is to produce crystals, the most common approach is to simply use all of the available samples in crystallization trials, as the screening *may* produce crystals, but other biophysical analyses certainly will not.

More recently, the DLS machines have become higher throughput, enabling the analysis of many samples in quick succession—this technique can now be used to compare side by side the same protein construct in different formulations, for example, allowing one to use the predictive power of the technique to select more likely candidates for further processing. This makes the technique more useful, as instead of just being a diagnostic tool it becomes a ranking tool.

Another, more recently adopted technology that can be used in much the same way is differential scanning fluorimetry (DSF). This technique, which uses an environmentally sensitive dye as a probe to measure protein unfolding, was initially proposed as an assay for small-molecule binding [36]. DSF is still used for screening libraries of small molecules [37] but is also used as a screen for testing formulation solutions in which the protein might be stable for crystallization and as a screen for additives that might help stabilize the protein [38, 39]. The general premise is that factors (formulations, pH, and small molecules) that stabilize the protein will increase the temperature at which it unfolds. As this assay is performed in 96- or 384-well plates in a standard RT-PCR machine, the assay has all the advantages of the high-throughput approach to DLS, but is generally much faster, with a normal DSF 96- or 384-well experiment taking an hour or less. There are two features of the DSF assay that give information about the protein sample—the temperature of the unfolding and the shape of the melt curve. The single value usually given for the DSF experiment is the melting temperature (T_m , or more correctly the temperature of hydrophobic exposure T_h) which is the inflection point of the unfolding curve.

Generally, the higher this number, the more stable the protein is (at the temperatures used for crystallization), and thus the more likely the sample is to be conformationally homogeneous, one of the requirements for crystallization. However, there is also a great deal of information in the shape of the fluorescence curve as well [40]—although the shape is harder to capture in a single number. The perfect curve shows a flat pre-transition and a steep melt transition. Less perfect curves show a downward curve in the pre-transition zone, suggesting that a subset of the sample is unfolded. In really bad melt curves, it is difficult or impossible to see a melting transition at all. This assay is potentially much more useful than the “traditional” single DLS measurement. The DSF experiment is still a low information containing experiment; it is recommended to run this in triplicate or even greater redundancy for two reasons—spurious measurements are easier to spot, and the replication gives an indication of the significance of the measurements—a significant shift in T_m has to be greater than the spread in the T_m for a set of identical or control measurements [40].

Some preliminary work from our lab suggests that DSF, like DLS, can be used as a predictor for crystallizability, confirming earlier reports [41]. In a recent experiment in our laboratory, a set of 60 proteins from *Streptococcus pneumoniae* were put through a very restricted crystallization screening strategy (96 conditions (JCSG+screen [42]), 1 temperature) and at the same time were put through a DSF formulation screen (“Buffer Screen 9” [40]). Of the 60 proteins, 3 appeared to be completely unfolded and gave no information about T_m in the DSF experiment: the experimentally derived curves had no structure and no indication of a melting step. These three proteins didn’t give anything remotely interesting in the crystallization screening either. For the remaining 57 samples, the average T_m was 47 °C for the proteins in their original buffer. More than half of the samples (34 of 57) produced well-shaped melt curves, and 11 of the proteins showed some indication of crystallizing—either large single crystals or unambiguous leads around which one could optimize. Of these 11 proteins, 6 had $T_m \geq 56$ °C (for a comparison, 13 of the set of 57 had $T_m \geq 56$ °C—see Fig. 1). It appears that a relatively high T_m of the protein in its current formulation may be a reasonable tool to predict crystallization success. There was a significant shift in the average T_m (47 °C to 52 °C) for the proteins in their “best” buffer as determined from this assay, and there were two formulations in which proteins showed a stability maximum—either slightly alkaline (HEPES or Tris) systems, pH 7.5–8, or slightly acid (citrate or Bis-tris) systems, pH 6.5. This suggests that swapping out the current “structural genomics” approach of one protein formulation for all proteins may be another relatively easy point at which to intervene to improve the crystallization success rate.

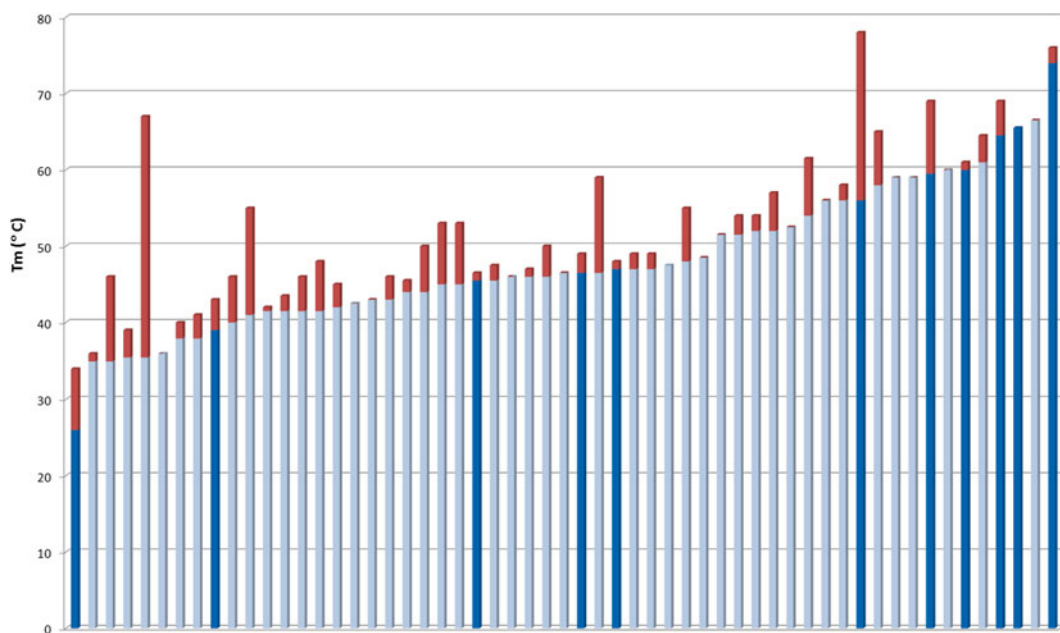


Fig. 1 Thermal shift assay. Of the 60 proteins in the *S. pneumoniae* sample set, 57 gave melt curves from which a T_m could be extracted. The gray/blue bars show the melt temperature of the samples in their initial buffer; the red bar shows the increase in T_m that could be obtained from another buffer/salt combination found in the Buffer Screen 9 protocol. Fifteen of the samples showed no improvement at all in any other buffer system, and 42 showed a preference for other formulations. The increase in T_m obtained in the other formulations ranged from 0.5 to 31.5 °C, with an average increase (over all 57 samples) of 4 °C. The bars in dark blue represent samples that yielded crystals. There is a statistically significant ($p=0.0182$) preference of proteins with a T_m of over 55 °C to crystallize

3 Crystallization Strategies

Structural genomics has contributed a great deal to the current state of crystallization: not only have these projects pushed the development of crystallization technology—such as imagers and low-volume dispensers—they have also introduced the concept of a process pipeline into the field [43]. Pre-structural genomics crystallization projects tended to involve two clear steps—a “screening” step, where factors that were positively correlated to crystallization were identified, and an “optimization” step, where the factors identified in the screening stage were recombined or otherwise tweaked until well-diffracting crystals were obtained. Structural genomics took a different approach: crystals are identified after a single, extensive set of crystallization trials. Although this approach is often called “screening” (“shotgun screening” [44, 45]), we choose to use the word “screying” as the goal is fundamentally different. “Screening” involves the identification of positive factors

which will be recombined in subsequent optimization experiments; “screaming” is an attempt to identify crystals for diffraction experiments. Not only are the goals of the two approaches different, but the initial crystallization experiments that are set up should vary, depending on the approach taken. For screening, the initial experiments should be chosen to identify all factors which are favorable for crystal formation: temperature, techniques, protein concentration as well as chemicals, and pH. The screaming approach would be better served by selecting only those crystallization conditions which have been shown to be most successful in the past [46].

4 Hotspots in Crystallization Space

Many of the conditions from the original Jancarik and Kim sparse matrix screen [21] (e.g., Crystal Screen from Hampton Research) are found unaltered as the crystallization conditions given in PDB. Is this a result of these being particularly successful crystallization conditions or simply a function of them being trialed more often? In order to answer the question “Are some crystallization conditions more likely to yield crystals than others?”, we need the success rate (i.e., successes compared to total number of trials). Although there is information about successful crystallization conditions available from public repositories (e.g., the PDB), the corresponding information about which conditions were trialed in order to produce those successful conditions is much harder to obtain. The basic concept of hotspots (i.e., successful conditions) in crystallization space was provided by structural genomics projects, where large numbers of protein constructs have been trialed against the same set of initial crystallization conditions. The results of these projects invariably show that a subset of the initial conditions will produce most of the crystals—for example, in work done at the Joint Center for Structural Genomics (JCSG), 480 commercial conditions were used to screen 542 proteins from *Thermotoga maritima*. An analysis of these crystallization experiments showed that 94 % of the crystals could have been obtained if only 192 of the conditions were used [43].

We can use the crystallization data in the PDB to suggest which of the commercial screens would be good starting points for either a screaming or a screening approach to crystallization. Screening campaigns will select screens which contain the chemical factors which are most commonly found in successful conditions and will contain conditions that combine these in simple binary combinations (emulating a balanced design of experiment (DOE) approach to testing factors), as the assumption cannot be made that a factor has only a positive or neutral effect on crystal formation. Screaming campaigns will select screens that look as similar as possible to the most successful conditions in the PDB.

5 Mining the PDB

The REMARK280 field of the PDB is probably the richest source of successful crystallization conditions that is available, and these data have been mined previously [46–50]. After the initial work of parsing and cleaning up the PDB crystallization information—a nontrivial task—we can find complete factors (strings that contain a *concentration value*, a *unit*, a *chemical name*, and (possibly) a *pH*) and conditions (sets of factors) in the REMARK280 field, and then we use the distance metric which underpins the C6 Web tool [51] to analyze the data. One of the confounding features of this analysis is the number of similar sequences in the PDB—there are over 300 PDB entries containing the hen egg white lysozyme sequence, and over 150 of these have an associated crystallization condition consisting of the very familiar combination of sodium acetate and sodium chloride. We need to do the analysis for the search for hotspot conditions on PDB data which has been purged of duplicate protein sequences and conditions (nonredundant PDB or NR-PDB). A set of the 96 most successful conditions as suggested by the REMARK280 of the NR-PDB is available as the Shotgun screen from the C6 Web tool (c6.csiro.au); the most successful conditions are shown in Table 1. The NR-PDB can also be used to find the most popular chemical factors in successful crystallization space, and a list of the top factors is shown in Table 2.

6 The (Over)abundance of Screens

There are at least 260 screens that are either available or have been available commercially [51]. Are there currently available crystallization kits that fulfill our requirements for initial trials for both the screening and scrying approaches to initial crystallization trials? We can test this using the C6 Web tool: comparing the PDB screen obtained from the crystallization information in the PDB with commercial screens picks up the ProPlex screen and Crystal Screen HT (from Molecular Dimensions and Hampton Research respectively, as well as the equivalent screens from other vendors) as being the best matches, suggesting that these screens are an appropriate start for a scrying strategy. We can look for screens that contain a large number of conditions containing the most likely factors for the screening approach; Table 2 shows the top chemicals found in the NR-PDB and the commercial screens that contain the highest number of conditions with these chemicals.

Table 1

The most abundant ten conditions in the NR-PDB (nonredundant Protein Data Bank) that match exactly to a commercial screen condition and the number of times the condition is found in the NR-PDB

Condition	Factor 1	Factor 2	Factor 3	Count in NR-PDB
1	30.000 w/v polyethylene glycol 4000	0.100 M tris buffer class pH 8.5	0.200 M magnesium chloride	95
2	2.000 M ammonium sulfate			85
3	20.000 w/v polyethylene glycol 3350	0.200 M acetate non-buffer class		80
4	2.000 M ammonium sulfate	0.100 M tris buffer class pH 8.5		80
5	20.000 w/v polyethylene glycol 3350	0.200 M citrate non-buffer class		79
6	20.000 w/v polyethylene glycol 4000	0.100 M HEPES buffer class pH 7.5	10.000 w/v 2-propanol	78
7	2.000 M ammonium sulfate	0.100 M HEPES buffer class pH 7.5	2.000 w/v polyethylene glycol 400	77
8	1.400 M citrate non-buffer class	0.100 M HEPES buffer class pH 7.5		75
9	30.000 w/v polyethylene glycol 4000	0.100 M tris buffer class pH 8.5	0.200 M acetate non-buffer class	74
10	30.000 w/v polyethylene glycol 4000	0.100 M tris buffer class pH 8.5	0.200 M lithium sulfate	64

The most abundant commercially available condition is found in at least ten commercial kits (Fig. 2)

Table 2

Chemicals found most often in the NR-PDB REMARK280 and the commercially available screens containing the most of those chemicals

Chemical name	Screens containing the chemical
Ammonium sulfate	The AmSO ₄ Suite (Qiagen)
Polyethylene glycol 3350	PEG/Ion HT Screen (Hampton Research)
Polyethylene glycol 4000	The PEGs II Suite (Qiagen)
Sodium chloride	MemStart and MemSys HT-96 (Molecular Dimensions)
Tris buffer class	Wizard Full (I and II) (Rigaku Reagents)

C6 Comparison of Crystallisation Conditions @ C3

Welcome to C6: janetn [Logout](#) [Update Password](#)

Select a report: [Reset Query](#)

Specify: [Switch](#) [Help](#)

Property: [Add new chemical](#) [Remove ticked chemical\(s\)](#) [Remove all](#) [Display Report](#)

Chemical	Conc.	Units	pH	Tick
Selected Chem: polyethylene glycol 3350	20	w/v		<input type="checkbox"/>
Selected Chem: sodium acetate	0.2	M		<input type="checkbox"/>

User's Condition:

Chemical	Concentration	Units	pH
polyethylene glycol 3350	20	w/v	
sodium acetate	0.2	M	

Results of Comparison:

Rank	Score(a ¹)	Well	Screen	Property	Owner
1.	0.0000	B1	JBScreen PACT++ 3		Jena Bioscience
2.	0.0000	E7	JBScreen PACT++ HTS		Jena Bioscience
3.	0.0000	A1,B5	JBScreen PEG/Salt 1		Jena Bioscience
3.	0.0000	A1,B5	JBScreen PEG/Salt 2		Jena Bioscience
4.	0.0000	A1,A11,C1,C11	JBScreen PEG/Salt HTS		Jena Bioscience
5.	0.0000	F5	MCSG 1		Microlytic
6.	0.0000	B2	MCSG 2		Microlytic
7.	0.0000	E7	PACT Premier		Molecular Dimensions
8.	0.0000	D2	PEG/ion 2		Hampton Research
9.	0.0000	C3-C6,G2	PEG/ion HT Screen		Hampton Research
10.	0.0000	D3-D6	PEG/ion Screen		Hampton Research

Fig. 2 Screenshot of the “Find condition in screens” report in the C6 Web tool. Here the condition entered is the most common commercial condition found in the NR-PDB. This analysis shows that there are at least ten commercial screens that contain this condition

7 Robotics in Crystallization

After almost 200 years of protein crystallization, we can learn a lot by looking at what have been the steps along the way to where we are now. A recent user survey at the two MX (macromolecular crystallography) beamlines of the Australian Synchrotron (David Aragão, personal communication) suggested that in about 50 % of cases, some sort of automation was being used for at least part of the crystallization work that leads to the crystals that were interrogated at the beamlines. Another way of looking at this is that automation is not being used in the preparation of half the potentially diffraction quality crystals and that manual techniques are still very valid in the preparation of well-diffracting protein crystals.

Why has crystallization automation not completely superseded laborious hand setups? Two reasons spring to mind: automation is not accessible to some in the community, and the processes that have been amenable to automation are not sufficient to produce

well-diffracting crystals. Consider the latter point—early trials with automation [27, 52] were based on the “structural genomics” concept—whereby a set of standard protocols would be used to produce purified protein, which would be fed into a crystallization process that was usually based on a number of commercially available screens. This approach is still being used, and results from the Protein Structure Initiative in the USA (PSI) suggest that of the ten purified proteins treated in this way, four will crystallize, and one will produce well-diffracting crystals [53]. Extrapolating further, simple scrying gives a 10 % success rate. Another estimation of the success of just using scrying can be garnered from the literature. Looking at more than 200 crystallization papers that were published in *Acta Crystallographica Section F* (a journal that focuses on crystallization and structure communications), over 75 % reported having to do one or more cycles of optimization to obtain the crystals reported in the paper [54]. There are a number of factors that can be adjusted in a crystallization campaign—temperature, drop size, drop ratio, the type of experiment that is set up, and the crystallization screens used, but generally these can be standardized: 20 °C, 200 nl, 1:1, sitting drop vapor diffusion would be very typical. It is clear from DSF experiments (and other reports in the literature [55–57]) that the formulation of the protein is a significant factor in crystallization success, but beyond the recognition that a general formulation should have some salt and some buffering chemical, the rationale behind the choice of “standard” formulation appears to be more driven by practical concerns (e.g., cost) rather than any widespread trials. We suspect that most of the formulations will be 20–50 mM HEPES or Tris buffer, with a pH between 7 and 8 and with 50–200 mM NaCl [54]. There is a balance between having a “soluble” protein—or at least one that stays in solution—which is more likely with higher concentrations of salt and/or other chemicals such as glycerol, and having a formulation which is dilute enough that the crystallization condition will perturb it enough to engender crystallization. There is little consensus on what initial trial conditions are best although without doubt it usually consists of some subset of the commercially available conditions [54] (rather than being collections of crystallization conditions developed individually in each laboratory). The main point of difference in the initial, automated screening strategies is simply in the number of trials that is set up [42, 58, 59].

One of the questions that comes up with the number of screening conditions is the question of coverage—are the extra trials covering the same area of crystallization space, or are they testing new parts of space—tools like the C6 give some measure of the area covered by a set of drops. Simply covering a wider swathe of possible crystallization space is not necessarily appropriate, as nucleation is a necessary first step towards a crystal and is a stochastic process; oversampling in popular areas of crystallization space may

increase the chances of success [59] more than sampling less successful crystallization conditions. Given this, the recent reemergence of seeding as a technique not just for optimization but for initial crystallization trials [60–63] is perhaps the most important recent trend in macromolecular crystallization. This process, which is called matrix seeding, uses any preliminary crystals as seeds in a syringing approach; although it generally uses seeds made from crystals of the same protein that one is trying to crystallize, cross matrix seeding has been shown to be successful as well, particularly in cases of antibodies (Fabs) and Fab/antigen complexes [64]. In this process a seeding solution is added as a third component of the crystallization droplet, so that the final volume of the seeding solution is about 10 % of the total volume of the drop. The success of syringing with seeding gives us some indication of how often nucleation is the limiting event in protein crystal growth, although other studies have also shown this [65]. The importance of seeding as part of the initial crystallization experiments is seen in the application of robotics to seeding—the Mosquito (TTP Labtech, UK) and Oryx (Douglas Instruments, UK) robots are well known to be appropriate for seeding [26, 61], but more recently the Gryphon device (Art Robbins Industries, USA) has used the syringe (initially designed for dispensing lipidic mesophase in membrane protein trials) as a dispenser for a seed stock.

8 Current State of the Art

Although there is always new instrumentation, new crystallization kits that are touted, and reworks of old or even new types of plastic consumables that become available, in essence the crystallization experiment has changed little in the last 40 years. The major difference is in the size of the droplets that are used—a quote from Alexander McPherson’s review published in 1976, “With this approach, a microdroplet of mother liquor (as small as 5 μ l) can be used...” [13], whereas today the more usual drop size is 1 μ l or less. Are there any indications that these tried and true approaches will change in the foreseeable future? The emergence of FELs (free electron laser) as a tool in macromolecular crystal structure determination may change the emphasis from larger crystals to many tiny crystals, but this type of crystallization offers its own challenges [66]. Many of the newer (improved!) crystallization consumables have been designed with the view to collecting X-ray diffraction data from crystals in their original growth containers, which in turn harks back to the past, where room temperature data was routinely collected.

Our initial analogy of crystallization to sequencing can now be answered in a different way. DNA sequencing is essentially a single-step process which has not only been automated, but multiple

different methods have now been developed to accomplish this step (next-generation sequencers). Crystallization on the other hand is a multistep process with many more variables (pH, temperature, chemical composition, and physical parameters) that determine whether a successful outcome is likely. The probability of crystal formation is low, making it a more difficult process to automate, and our attempts at automation have focused only on certain steps of the process. And in crystallization, the different methods may very well give different outcomes (crystals vs. no crystals or one morphology vs. another).

Additionally, the recent focus (at least for those labs with automation) has been on scribing and not screening. To improve our hit rates in crystallization for diffraction quality crystals, it seems that more effort in optimization may well improve our ability to obtain quality structures. For this, and in the absence of the next big thing which will change the field, we look to the past and learn that what are new today—better characterization of the sample, seeding, and using past conditions in order to maximize success—are just echoes of the past.

References

1. Halliburton W (1887) Memoirs: on the haemoglobin crystals of rodents' blood. *Q J Microsc Sci* 2:181–199
2. Sumner JB (1926) The isolation and crystallization of the enzyme urease preliminary paper. *J Biol Chem* 69:435–441
3. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The protein data bank: a computer-based archival file for macromolecular structures. *J Mol Biol* 112:535–542
4. Jou WM, Haegeman G, Ysebaert M, Fiers W (1972) Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* 237:82–88
5. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison CA, Slocumbe PM, Smith M (1977) Nucleotide sequence of bacteriophage ϕ X174 DNA. *Nature* 265:687–695
6. Venter JC (2001) The sequence of the human genome. *Science* 291:1304–1351
7. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921
8. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci* 74:5463–5467
9. Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* 242:84–89
10. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* 18:630–634
11. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* [online]. <http://www.nature.com/doifinder/10.1038/nature03959>. Accessed 1 Jan 2014
12. Bernal JD, Crowfoot D (1934) X-ray photographs of crystalline pepsin. *Nature* 133:794–795

13. McPherson A Jr (1976) The growth and preliminary investigation of protein and nucleic acid crystals for X-ray diffraction analysis. *Methods Biochem Anal* 23:249–345
14. Janssen FW, Ruelius HW (1968) Alcohol oxidase, a flavoprotein from several Basidiomycetes species: crystallization by fractional precipitation with polyethylene glycol. *Biochim Biophys Acta* 151:330–342
15. McPherson A (1976) Crystallization of proteins from polyethylene glycol. *J Biol Chem* 251:6300–6303
16. Emtage JS, Angal S, Doel MT, Harris TJ, Jenkins B, Lilley G, Lowe PA (1983) Synthesis of calf prochymosin (prorennin) in *Escherichia coli*. *Proc Natl Acad Sci* 80:3671–3675
17. Moks T, Abrahmsén L, Österlöf B, Josephson S, Östling M, Enfors S-O, Persson I, Nilsson B, Uhlén M (1987) Large-scale affinity purification of human insulin-like growth factor I from culture medium of *Escherichia coli*. *Biotechnology* 5:379–382
18. Germino J, Bastia D (1984) Rapid purification of a cloned gene product by genetic fusion and site-specific proteolysis. *Proc Natl Acad Sci* 81:4692–4696
19. Hochuli E, Bannwarth W, Döbeli H, Gentz R, Stüber D (1988) Genetic approach to facilitate purification of recombinant proteins with a novel metal chelate adsorbent. *Biotechnology* 6:1321–1325
20. Hopp TP, Prickett KS, Price VL, Libby RT, March CJ, Pat Cerretti D, Urdal DL, Conlon PJ (1988) A short polypeptide marker sequence useful for recombinant protein identification and purification. *Biotechnology* 6:1204–1210
21. Jancarik J, Kim SH (1991) Sparse matrix sampling: a screening method for crystallization of proteins. *J Appl Crystallogr* 24:409–411
22. Hope H (1988) Cryocrystallography of biological macromolecules: a generally applicable method. *Acta Crystallogr B* 44:22–26
23. Parkin S, Hope H (1998) Macromolecular cryocrystallography: cooling, mounting, storage and transportation of crystals. *J Appl Crystallogr* 31:945–953
24. Bard J, Ercolani K, Svenson K, Olland A, Somers W (2004) Automated systems for protein crystallization. *Methods* 34:329–347
25. Hiraki M, Kato R, Nagai M, Satoh T, Hirano S, Ihara K, Kudo N, Nagae M, Kobayashi M, Inoue M, Uejima T, Oda S, Chavas LMG, Akutsu M, Yamada Y, Kawasaki M, Matsugaki N, Igarashi N, Suzuki M, Wakatsuki S (2006) Development of an automated large-scale protein-crystallization and monitoring system for high-throughput protein-structure analyses. *Acta Crystallogr D Biol Crystallogr* 62:1058–1065
26. Newman J, Pham TM, Peat TS (2008) Phoenix experiments: combining the strengths of commercial crystallization automation. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 64:991–996
27. Walter TS, Diprose J, Brown J, Pickford M, Owens RJ, Stuart DI, Harlos K (2003) A procedure for setting up high-throughput nanolitre crystallization experiments. I. Protocol design and validation. *J Appl Crystallogr* 36:308–314
28. Flot D, Mairs T, Giraud T, Guijarro M, Lesourd M, Rey V, van Brussel D, Morawe C, Borel C, Hignette O, Chavanne J, Nurizzo D, McSweeney S, Mitchell E (2009) The ID23-2 structural biology microfocus beamline at the ESRF. *J Synchrotron Radiat* 17:107–118
29. Slabinski L, Jaroszewski L, Rodrigues APC, Rychlewski L, Wilson IA, Lesley SA, Godzik A (2007) The challenge of protein structure determination—lessons from structural genomics. *Protein Sci* 16:2472–2482
30. Jahandideh S, Jaroszewski L, Godzik A (2014) Improving the chances of successful protein structure determination with a random forest classifier. *Acta Crystallogr D Biol Crystallogr* 70:627–635
31. Dale GE, Oefner C, D'Arcy A (2003) The protein as a variable in protein crystallization. *J Struct Biol* 142:88–97
32. Sagemark J, Kraulis P, Weigelt J (2010) A software tool to accelerate design of protein constructs for recombinant expression. *Protein Expr Purif* 72:175–178
33. Lorber B, Fischer F, Bailly M, Roy H, Kern D (2012) Protein analysis by dynamic light scattering: methods and techniques for students. *Biochem Mol Biol Educ* 40:372–382
34. Ferré-D'Amaré AR, Burley S (1994) Use of dynamic light scattering to assess crystallizability of macromolecules and macromolecular assemblies. *Structure* 2:357–359
35. Ferré-D'Amaré AR, Burley SK (1997). In: *Macromolecular Crystallography Part A Methods in Enzymology*, pp 157–166 [online]. <http://www.sciencedirect.com/science/article/pii/S0076687997760567>. Accessed 24 Feb 2013
36. Pantoliano MW, Petrella EC, Kwasnoski JD, Lobanov VS, Myslik J, Graf E, Carver T, Asel E, Springer BA, Lane P, Salemme FR (2001) High-density miniaturized thermal shift assays as a general strategy for drug discovery. *J Biomol Screen* 6:429–440

37. McMahon RM, Scanlon MJ, Martin JL (2013) Interrogating fragments using a protein thermal shift assay. *Aust J Chem* 66:1502
38. Niesen FH, Berglund H, Vedadi M (2007) The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nat Protoc* 2:2212–2221
39. Vedadi M, Arrowsmith CH, Allali-Hassani A, Senisterra G, Wasney GA (2010) Biophysical characterization of recombinant proteins: a key to higher structural genomics success. *J Struct Biol* 172:107–119
40. Seabrook SA, Newman J (2013) High-throughput thermal scanning for protein stability: making a good technique more robust. *ACS Comb Sci* 15:387–392
41. Dupeux F, Röwer M, Seroul G, Blot D, Márquez JA (2011) A thermal stability assay can help to estimate the crystallization likelihood of biological samples. *Acta Crystallogr D Biol Crystallogr* 67:915–919
42. Newman J, Egan D, Walter TS, Meged R, Berry I, Ben Jelloul M, Sussman JL, Stuart DI, Perrakis A (2005) Towards rationalization of crystallization screening for small- to medium-sized academic laboratories: the PACT/JCSG+ strategy. *Acta Crystallogr D Biol Crystallogr* 61:1426–1431
43. Lesley SA, Kuhn P, Godzik A, Deacon AM, Mathews I, Kreusch A, Spraggon G, Klock HE, McMullan D, Shin T et al (2002) Structural genomics of the *Thermotoga maritima* proteome implemented in a high-throughput structure determination pipeline. *Proc Natl Acad Sci U S A* 99:11664
44. McPherson A, Gavira JA (2013) Introduction to protein crystallization. *Acta Crystallogr Sect F Struct Biol Commun* 70:2–20
45. Luft JR, Newman J, Snell EH (2014) *Acta Crystallogr Sect F Struct Biol Commun*. 70:835–853
46. Fazio VJ, Peat TS, Newman J (2014) *Acta Crystallogr Sect F Struct Biol Commun* 70:1303–1311
47. Tung M, Gallagher DT (2008) The biomolecular crystallization database version 4: expanded content and new features. *Acta Crystallogr D Biol Crystallogr* 65:18–23
48. Peat TS, Christopher JA, Newman J (2005) Tapping the Protein Data Bank for crystallization information. *Acta Crystallogr D Biol Crystallogr* 61:1662–1669
49. Gorrec F (2009) The MORPHEUS protein crystallization screen. *J Appl Crystallogr* 42:1035–1042
50. Newstead S, Ferrandon S, Iwata S (2008) Rationalizing α -helical membrane protein crystallization. *Protein Sci* 17:466–472
51. Newman J, Fazio VJ, Lawson B, Peat TS (2010) The C6 Web tool: a resource for the rational selection of crystallization conditions. *Cryst Growth Des* 10:2785–2792
52. Page R, Grzechnik SK, Canaves JM, Spraggon G, Kreusch A, Kuhn P, Stevens RC, Lesley SA (2003) Shotgun crystallization strategy for structural genomics: an optimized two-tiered crystallization screen against the *Thermotoga maritima* proteome. *Acta Crystallogr D Biol Crystallogr* 59:1028–1037
53. Newman J, Bolton EE, Müller-Dieckmann J, Fazio VJ, Gallagher DT, Lovell D, Luft JR, Peat TS, Ratcliffe D, Sayle RA, Snell EH, Taylor K, Vallotton P, Velanker S, von Delft F (2012) On the need for an international effort to capture, share and use crystallization screening data. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 68:253–258
54. Newman J, Burton DR, Caria S, Desbois S, Gee CL, Fazio VJ, Kvanakul M, Marshall B, Mills G, Richter V, Seabrook SA, Wu M, Peat TS (2013) Crystallization reports are the backbone of *Acta Cryst. F*, but do they have any spine? *Acta Crystallogr Sect F Struct Biol Cryst Commun* 69:712–718
55. Zhang C-Y, Wu Z-Q, Yin D-C, Zhou B-R, Guo Y-Z, Lu H-M, Zhou R-B, Shang P (2013) A strategy for selecting the pH of protein solutions to enhance crystallization. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 69:821–826
56. Jancarik J, Pufan R, Hong C, Kim SH, Kim R (2004) Optimum solubility (OS) screening: an efficient method to optimize buffer conditions for homogeneity and crystallization of proteins. *Acta Crystallogr D Biol Crystallogr* 60:1670–1673
57. Benvenuti M, Mangani S (2007) Crystallization of soluble proteins in vapor diffusion for x-ray crystallography. *Nat Protoc* 2:1633–1651
58. Berry IM, Dym O, Esnouf RM, Harlos K, Meged R, Perrakis A, Sussman JL, Walter TS, Wilson J, Messerschmidt A (2006) SPINE high-throughput crystallization, crystal imaging and recognition techniques: current state, performance analysis, new technologies and future aspects. *Acta Crystallogr D Biol Crystallogr* 62:1137–1149
59. Gorrec F (2013) The current approach to initial crystallization screening of proteins is under-sampled. *J Appl Crystallogr* 46:795–797

60. Till M, Robson A, Byrne MJ, Nair AV, Kolek SA, Shaw Stewart PD, Race PR (2013) Improving the success rate of protein crystallization by random microseed matrix screening. *J Vis Exp* [online]. <http://www.jove.com/video/50548/improving-success-rate-protein-crystallization-random-microseed>. Accessed 1 Jan 2014
61. D'Arcy A, Villard F, Marsh M (2007) An automated microseed matrix-screening method for protein crystallization. *Acta Crystallogr D Biol Crystallogr* 63:550–554
62. Ireton GC, Stoddard BL (2004) Microseed matrix screening to improve crystals of yeast cytosine deaminase. *Acta Crystallogr D Biol Crystallogr* 60:601–605
63. D'Arcy A, Bergfors T, Cowan-Jacob SW, Marsh M (2014) *Acta Crystallogr Sect F Struct Biol Commun* 70:1117–1126
64. Obmolova G, Malia TJ, Teplyakov A, Sweet R, Gilliland GL (2010) Promoting crystallization of antibody–antigen complexes *via* microseed matrix screening. *Acta Crystallogr D Biol Crystallogr* 66:927–933
65. Newman J, Xu J, Willis MC (2007) Initial evaluations of the reproducibility of vapor-diffusion crystallization. *Acta Crystallogr D Biol Crystallogr* 63:826–832
66. Schlichting I, Miao J (2012) Emerging opportunities in structural biology with X-ray free-electron lasers. *Curr Opin Struct Biol* 22:613–626

Part III

Membrane Proteins

Screening of Stable G-Protein-Coupled Receptor Variants in *Saccharomyces cerevisiae*

Mitsunori Shiroishi and Takuya Kobayashi

Abstract

G-protein-coupled receptors (GPCRs) are not only the largest protein family, but as a whole, they represent the largest group of therapeutic drug targets. Recent successes in the determination of GPCR structures have relied on the stabilization of receptors to overcome the difficulties in expression and purification. Although a large quantity of purified protein is needed for structural determination, the majority of wild-type GPCRs are too unstable to express and purify on a large scale. Therefore, rapid screening of highly expressed stable receptor “variants” is crucial. It has been demonstrated that fusing green fluorescent protein (GFP) to a target membrane protein facilitates the evaluation of the physical properties of the membrane protein in detergent. Furthermore, the budding yeast *Saccharomyces cerevisiae* enables rapid construction of an expression vector via its own efficient homologous recombination system. Herein, we describe the protocols for rapid construction and screening of stable GPCR variants using GFP and *S. cerevisiae*.

Key words G-protein-coupled receptor (GPCR), *Saccharomyces cerevisiae*, Green fluorescent protein (GFP), Protein engineering, Structural biology

1 Introduction

G-protein-coupled receptors (GPCRs) are the largest protein family with more than 800 genes coded in the human genome [1]. Many GPCRs play important roles in signal transduction events throughout the body. Therefore, they are major therapeutic drug targets and represent more than 30 % of the market share of all prescription drugs [2]. The high-resolution crystal structures of the target receptors provide good initial models for structure-based approaches to drug screening and drug design and help to accelerate drug discovery. Recent advances in the methods for the preparation and crystallization of recombinant receptors, together with advances in X-ray data collection systems, have enabled structural analysis of GPCRs. However, only a limited number of GPCRs have been successfully expressed and purified sufficiently for these analyses.

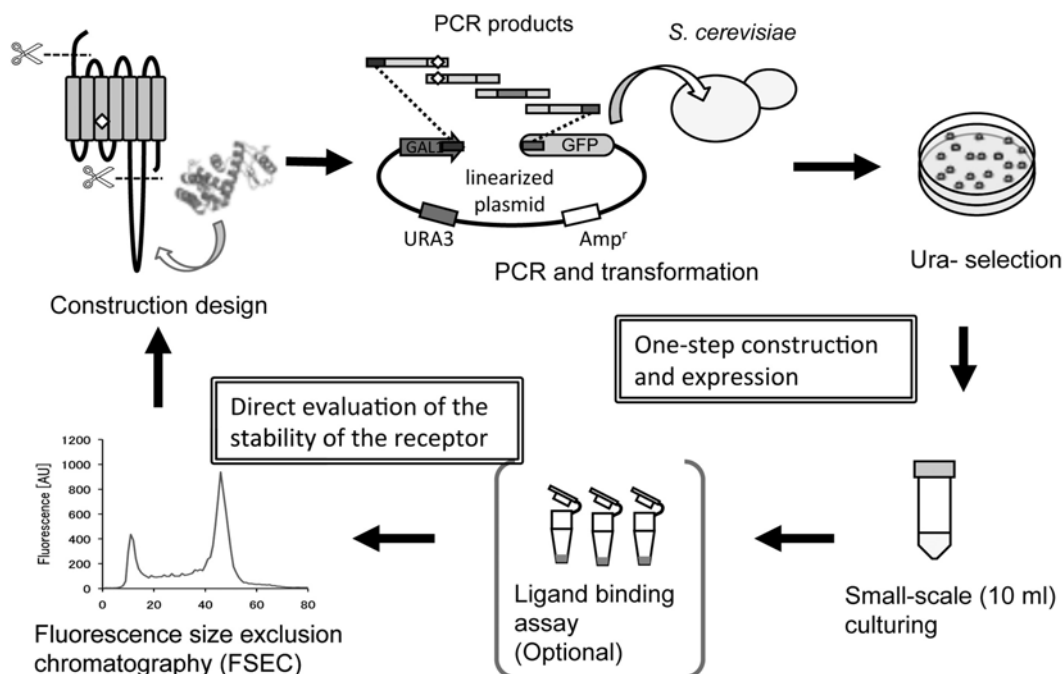


Fig. 1 Overview of the screening platform of GPCR variants in *S. cerevisiae*. *S. cerevisiae* enables one-step construction and expression of the target receptor variants from the PCR fragments and the plasmid, thus eliminating the vector construction step in *E. coli*. FSEC enables direct evaluation of the stability of the receptors solubilized in detergent without purification. A ligand binding assay using a radiolabeled ligand provides information on the functionality of the expressed receptor. This is optional, since accessibility to a radioisotope facility and availability of the appropriate radiolabeled ligand may depend on the research institution and the receptor

For structural, biochemical, and biophysical studies of GPCRs, a large amount of highly purified sample is required. However, mammalian membrane proteins are often unstable and expressed in heterologous hosts at a level that is too low for the purification of sufficient protein. One solution to this bottleneck is to construct a stable and highly expressed receptor variant. Because this is a process of trial and error, a screening system is needed.

We employed a screening platform using *S. cerevisiae* and green fluorescent protein (GFP) for the rapid construction and evaluation of GPCR variants. The overall process is described in Fig. 1. This platform is based on the system developed by Drew et al. [3]. With the help of the high recombination efficiency of *S. cerevisiae*, construction of the expression vector is accomplished in one step in a yeast cell by co-transformation of a linearized plasmid and DNA fragments (PCR products) of the target receptor. The plasmid used here contains the GFP coding region immediately downstream of the insertion site for the target receptor, so that the receptor is expressed as a C-terminal GFP fusion. Hence, we can directly evaluate the expression level and the physical stability of the solubilized receptor without purification by following its fluorescence [4].

This platform enables us to perform a screening cycle within 6–7 days, which is much faster than the use of the methylotrophic yeast *Pichia pastoris* (16–18 days) and insect cells (30–35 days) [5].

The stabilized receptor variants evaluated in *S. cerevisiae* also showed better characteristics than wild-type receptors in other expression hosts, such as *P. pastoris* or insect cells. Using this platform, we could identify stable human histamine H₁ receptor (hH₁R) variants. The expression level of the stabilized hH₁R variant was increased to 65 pmol/mg compared with negligible expression of the wild-type receptor at first screening in *S. cerevisiae*, which is enough for purification and structural studies. Using the variant, we determined the structure of hH₁R complexed with the inhibitor doxepin [6]. In this chapter, we describe the practical protocols of the screening platform in *S. cerevisiae* for producing stable GPCR variants. This platform is applicable to the expression and purification of other integral membrane proteins.

2 Materials

2.1 Transformation of *S. cerevisiae*

1. Thermal cycler.
2. Centrifuge with swinging bucket rotor and holders for 15 and 50 ml Falcon tubes (*see Note 1*).
3. Sterile 50 ml plastic tubes with aerated cap, e.g., TPP bioreactor tube (Techno Plastic Products, Switzerland). Alternatively, sterilized 24×200 mm glass test tubes with aluminum caps can be used.
4. A high-fidelity DNA polymerase with its associated specific reaction buffer and dNTP mix.
5. *S. cerevisiae* strain FGY217 (*MAT α* , *ura3-52*, *lys2 Δ 201*, *pep4 Δ*) [7].
6. Plasmid pDDGFP-2 [3, 8] which is derived from the vector pRS426GAL1 and carries a TEV protease recognition site, GFP, and octa-histidine tag sequence.
7. 50 % (w/v) D-(+)-glucose. Filter sterilize and store at 4 °C.
8. Yeast extract peptone dextrose (YPD) medium: 1 % yeast extract, 2 % Bacto peptone, and 2 % glucose (dextrose). Add 2 % agar for YPD plates. Glucose should be added after autoclaving. Store at 4 °C.
9. 1 M lithium acetate. Autoclave and store at 4 °C.
10. 0.1 M lithium acetate. Dilute 1 M lithium acetate with pure water. Autoclave and store at 4 °C.
11. 50 % (w/v) polyethylene glycol 3,350. Filter sterilize and store at 4 °C.
12. Carrier DNA solution: 2 mg/ml deoxyribonucleic acid sodium salt from salmon testes in TE buffer. Store at –20 °C.

13. Synthetic complete (SC) medium plate without uracil: 1.7 g yeast nitrogen base (without amino acids) and ammonium sulfate, 5 g ammonium sulfate, 1.92 g yeast synthetic dropout supplement (without uracil), 2 % glucose, and 20 g agar. Glucose should be added after autoclaving.

2.2 Small-Scale Expression and Membrane Preparation

1. Centrifuge with swing bucket rotor and holders for 15 and 50 ml Falcon tubes.
2. Fluorescence microplate reader.
3. Vortex-Genie 2 mixer with Turbomix attachment (Scientific Industries, Inc., Bohemia, NY).
4. Benchtop ultracentrifuge Optima TL100 (Beckman Coulter, Indianapolis) with a TLA-100.3 rotor equipped with a 1.5 ml tube adaptor.
5. Phase-contrast microscope ($\times 200$ – 400 magnification).
6. Sterile 50 ml plastic tube with aerated cap or sterilized 24×200 mm glass test tube.
7. 2.0 ml microfuge tube.
8. 1.5 ml microfuge tube for ultracentrifugation.
9. 25 % (w/v) galactose. Filter sterilize and store at 4 °C.
10. 0.5 mm glass beads (*see Note 2*).
11. SC medium (without uracil) and either 2 % glucose (for pre-culture) or 0.1 % glucose (for expression culture). Glucose should be added after autoclaving.
12. Nunc 96-well black microplate.
13. EDTA-free complete protease inhibitor cocktail.
14. Yeast suspension buffer: 50 mM Tris-HCl (pH 7.5), 5 mM EDTA, 10 % glycerol, 0.12 M sorbitol, and 1 \times complete EDTA-free protease inhibitor.
15. Membrane buffer: 50 mM Tris-HCl (pH 8.0), 120 mM NaCl, 20 % glycerol, and protease inhibitor cocktail.

2.3 Evaluation of the GPCR Variants by FSEC

1. High-performance liquid chromatography (HPLC) system or fast protein liquid chromatography (FPLC) system (such as AKTA series (GE Healthcare, United Kingdom) or BioLogic DuoFlow (BioRad, Hercules, CA)).
2. Fluorescence detector (*see Note 3*).
3. Benchtop ultracentrifuge with a rotor equipped with a 1.5 ml tube adaptor.
4. 1.5 ml microfuge tube for ultracentrifugation.
5. High-performance gel filtration column. Superose 6 10/300 (GE Healthcare) or Superdex 200 10/300 GL (GE

Healthcare). (The smaller column Superdex 200 5/150 enables more rapid analysis.)

6. *n*-Dodecyl- β -D-maltopyranoside (DDM) (*see* **Note 4**).
7. Cholesteryl hemisuccinate (CHS).
8. Gel filtration buffer: 20 mM Tris-HCl (pH 7.5), 150 mM NaCl, and 0.03 % DDM/0.006 % CHS.
9. Solubilization buffer: 50 mM Tris-HCl (pH 7.5), 200 mM NaCl, 1 % DDM/0.2 % CHS, and protease inhibitor cocktail.

2.4 Isolation of Plasmid from *S. cerevisiae* and Analysis of the DNA Sequence

1. Plasmid purification kit (e.g., QIAprep Miniprep Kit).
2. Vortex mixer.
3. 425–600 μ m acid-washed glass beads (Sigma-Aldrich).
4. Zymolyase 20T (Nakalai Tesque, Japan).
5. Sorbitol buffer: 1 M sorbitol and 0.1 M EDTA (pH 7.5).

3 Methods

3.1 Transformation of *S. cerevisiae*

1. Design receptor variants and the primers for PCR (*see* **Note 5**). We show an example of the primer design for a GPCR variant (Fig. 2a, b). Perform PCR using a high-fidelity polymerase. Analyze the products by agarose gel electrophoresis. No need for purification of PCR products.
2. Digest the plasmid pDDGFP-2 with the SmaI restriction enzyme. After digestion, incubate at 60 °C for 20 min to inactivate the enzyme. Dilute the reaction mixture by adding sterilized water to a plasmid concentration of ~40 ng/ μ l. No need for purification of the digested plasmid. Make aliquots and store at –20 °C. The digested plasmid can be stored for up to 1 year or more. We confirm that the plasmid works even after ten freeze–thaw cycles.
3. Streak *S. cerevisiae* strain FGY217 onto a YPD plate from a glycerol stock. Incubate the plate at 30 °C for 2–3 days. Once the colonies have grown, store the plate at 4 °C. The plate can be stored for up to 2 months.
4. Inoculate 5 ml YPD medium with a single colony of FGY217 from the YPD plate. Grow overnight at 30 °C in a shaking incubator at 300 rpm.
5. The next morning, check the OD₆₀₀ of the culture. Normally, it reaches 8–10. This culture can be stored at 4 °C for several days.
6. Add 50 ml of the YPD medium to a sterilized 200 ml baffled flask. Inoculate the medium with the culture so that the final OD₆₀₀ is 0.12. Grow at 30 °C in the shaking incubator until the OD₆₀₀ reaches 0.6–0.8. Usually, it takes 5–6 h. 50 ml of culture is enough for ~20 samples.

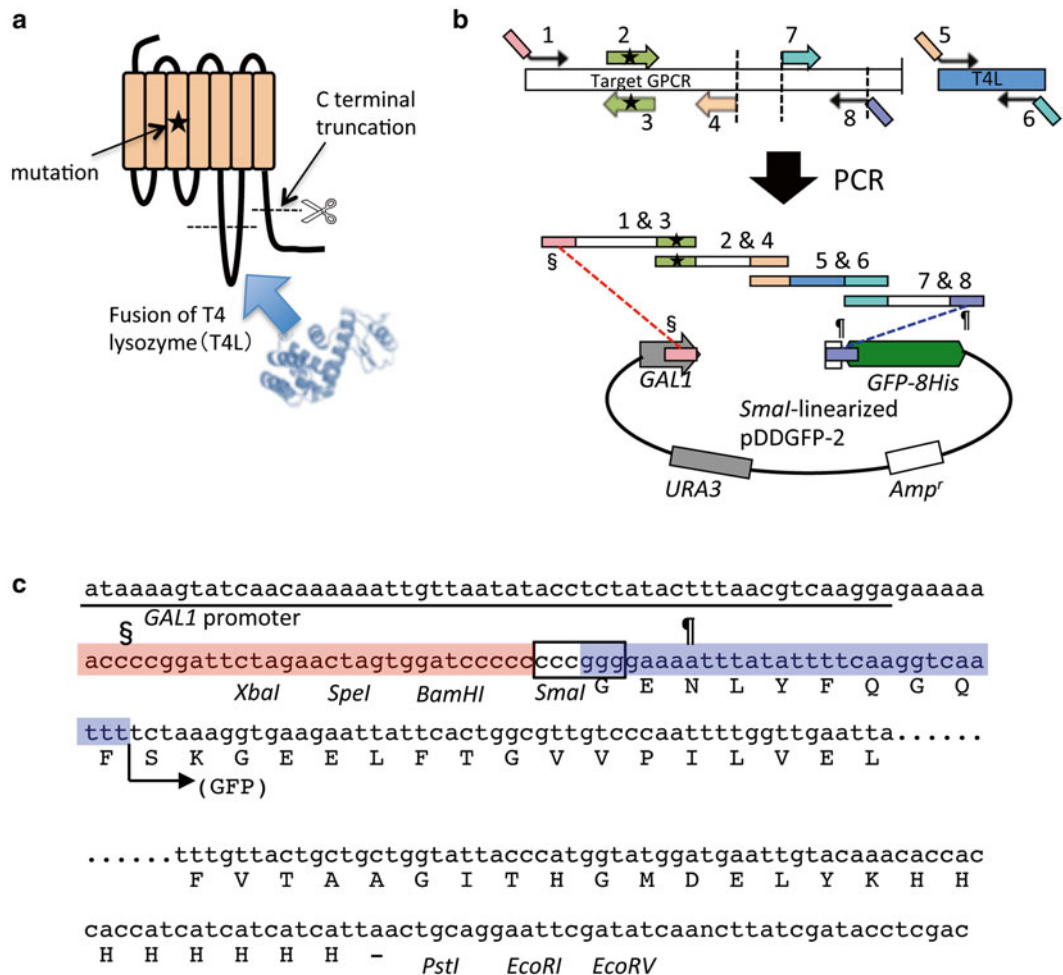


Fig. 2 Design of primers for producing variants in *S. cerevisiae*. (a) As an example, a GPCR variant with truncation of the N-terminal region, a mutation at the 341 position in TM3 and T4L fusion to the i3-loop is shown. (b) Eight primers [1–8] needed for producing the variant shown in (a). The four PCR fragments are generated using the indicated primer pairs with the full-length GPCR and T4L as templates. A total of 30 base-pair overlapping regions are necessary for homologous recombination in *S. cerevisiae*. In particular, the regions indicated by § and ¶ are required for integration into the plasmid. (c) The DNA sequence from the end of the *GAL1* promoter to the end of GFP–His8 in pDDGFP-2. Both shaded sequences (§ and ¶) correspond to the regions in (b)

7. Harvest the cells in a sterilized 50 ml conical tube at $3,000\times g$ for 5 min and decant off the supernatant. Resuspend the cell pellet in 20 ml sterile water.
8. Centrifuge the tube at $3,000\times g$ for 5 min and decant off the supernatant. Centrifuge again for 2 min, and discard the supernatant thoroughly using a pipette.
9. Resuspend the cell pellet in 1 ml 0.1 M LiAc, and transfer to a sterile 1.5 ml tube.
10. Centrifuge at $12,000\times g$ for 5 s, and remove the supernatant.

11. Resuspend the cell pellet in 950 μ l 0.1 M LiAc (total cell suspension is about 1 ml).
12. Add the following materials in order to sterile 1.5 ml tubes, and mix well by vortexing:
 - 1 μ l SmaI-cut pDDGFP-2 (~40 ng/ μ l).
 - 3 μ l of each PCR product (*see Note 6*).
 - X μ l sterile water (make volume up to 50 μ l at this step).
 - 25 μ l single-stranded carrier DNA solution.
 - 240 μ l 50 % PEG 3,350 solution.
 - 36 μ l 1.0 M LiAc.
13. Add 50 μ l cell suspension into the tube, and mix well by vortexing.
14. Incubate at 30 °C for 30 min.
15. Incubate at 42 °C for 20–25 min (heat shock).
16. Centrifuge at 8,000 $\times g$ for 15 s, and discard the supernatant.
17. Suspend the cell pellet in 200 μ l sterile water. Spread 50 μ l cell suspension onto a SC medium plate without uracil. The remaining cell suspension can be stored for several days.
18. Incubate at 30 °C for 2–3 days until the diameter of the colonies becomes more than 1 mm (*see Note 7*). Then store the plate at 4 °C. The plate can be stored for up to 2 months.

3.2 Small-Scale Expression and Membrane Preparation

1. Inoculate 3 ml SC medium containing 2 % glucose but no uracil, with a single colony of the yeast transformants. Grow overnight at 30 °C in the shaking incubator (*see Note 8*). The overnight culture can be stored at 4 °C for several days.
2. The following day, inoculate 10 ml SC medium containing 0.1 % glucose but no uracil with the overnight culture that is at an OD₆₀₀ of 0.12 (*see Note 9*).
3. Grow at 30 °C in the shaking incubator for 6–7 h until the OD₆₀₀ reaches around 0.6.
4. Add 850 μ l of 25 % galactose to a final concentration of 2 %, and grow at 30 °C in the shaking incubator for 20–22 h (*see Note 10*).
5. Transfer the cell culture to a 15 ml conical tube, and centrifuge at 3,000 $\times g$ for 5 min. Decant off the supernatant. Centrifuge at 3,000 $\times g$ for 1 min, and aspirate off the supernatant using a pipette, making sure to remove all the media.
6. Resuspend the cell pellet in 700 μ l yeast suspension buffer and transfer to a 2 ml tube. Keep the sample on ice for the remainder of the procedure.
7. Take 20 μ l of the cell suspension, and dilute with 180 μ l yeast suspension buffer. Mix well, and transfer the suspension into a

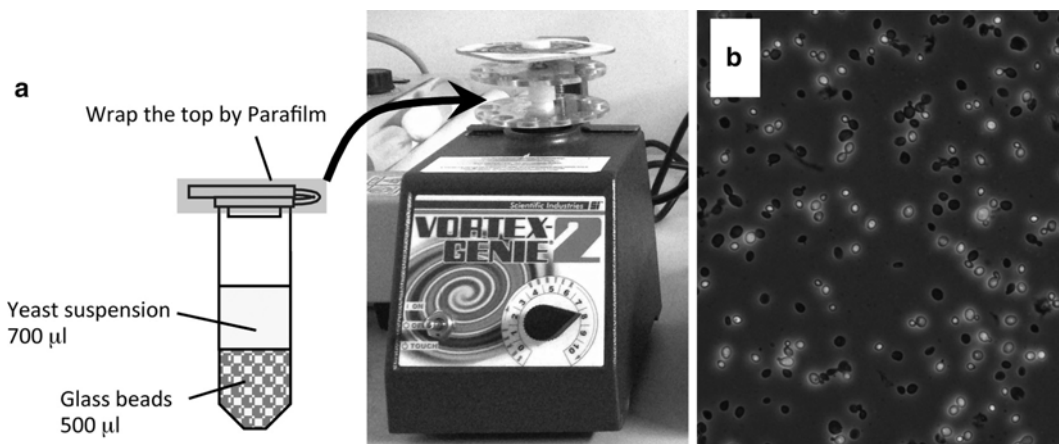


Fig. 3 Disruption of yeast cells. **(a)** Depiction of the yeast suspension in a 2 ml tube with glass beads and the cell disruptor (Vortex-Genie 2 mixer with Turbomix attachment). **(b)** Phase-contrast microscopic view of the disrupted *S. cerevisiae* cells. Disrupted cells appear black, and undisturbed cells appear white with a glow

96-well black microplate. Measure the GFP fluorescence at 490 nm excitation and 525 nm emission using the 515 nm filter (*see Note 11*).

8. Put 500 µl of 0.5 mm dry glass beads into the cell suspension in a 2 ml tube. Wrap the top of the tube with parafilm.
9. Disrupt the cells by vortexing at maximum speed for 10 min in the cold room. Check the cells with a phase-contrast microscope. More than 80 % of cells are expected to be disrupted (Fig. 3).
10. Centrifuge at $12,000\times g$ for 1 min at 4 °C to spin down unbroken cells and cell debris.
11. Transfer 500 µl of the supernatant into a 1.5 ml ultracentrifugation microtube. Take care not to transfer the glass beads.
12. Centrifuge at $100,000\times g$ for 30 min at 4 °C to collect the membrane. Discard the supernatant.
13. Add 50 µl membrane buffer and resuspend the membrane pellet using a pellet pestle. The membrane suspension can be stored on ice for 1 day. For long-term storage, freeze in liquid nitrogen and store at -80 °C.

3.3 Evaluation of GPCR Variants by FSEC (See Note 12)

1. Turn on the HPLC (or FPLC) system and fluorescence detector.
2. Equilibrate the gel filtration column with the gel filtration buffer.
3. Add 200 µl solubilization buffer to 50 µl of membrane suspension and solubilize the membrane by mild agitation for 1 h at 4 °C.
4. Centrifuge at $100,000\times g$ for 30 min at 4 °C.

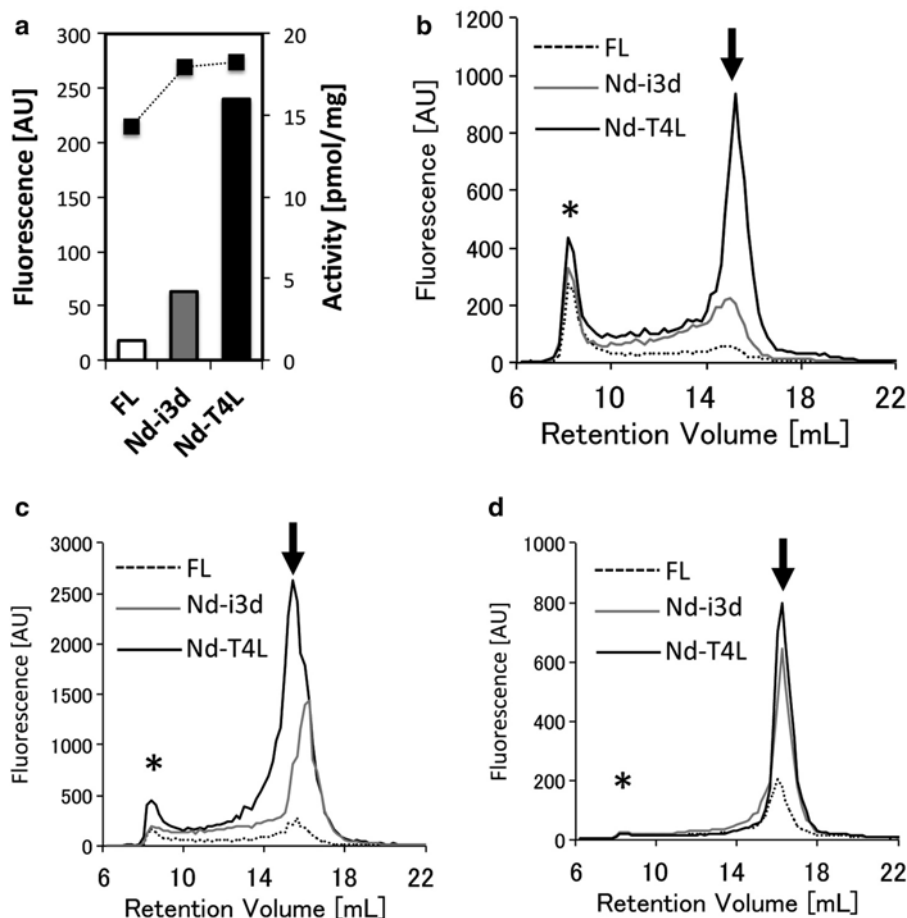


Fig. 4 Examples of the evaluation of GPCR variants. The provided examples show the results for the human histamine H_1 receptor. (a) Total expression level estimated from GFP fluorescence (black squares) and specific binding activity of the receptor variants expressed in *S. cerevisiae* for the radiolabeled antagonist [3H]-pyrilamine (bar graph). Full-length receptor (FL), the variant with N-terminal truncation and deletion of the i3-loop (Nd-i3d), and the variant with N-terminal truncation and fusion of T4L instead of the i3-loop (Nd-T4L). (b–d) The FSEC profiles of the receptor variants expressed in *S. cerevisiae* (b), *P. pastoris* (c), and Sf9 insect cells (d). FSEC was performed using a Superose 6 10/300 column. The asterisk indicates the void peak. The arrow indicates the peak of the monomeric receptor. Stabilized receptor variants in *S. cerevisiae* showed improved characteristics compared with wild-type receptors in the other hosts. In the case of H_1R , there is a good correlation between FSEC profiles and ligand bindings

5. Transfer the supernatant into a new 1.5 ml tube.
6. Apply 100–200 μ l of the supernatant to a Superose 6 10/300 or Superdex 200 10/300 column at a flow rate of 0.5 ml/min (Fig. 4b). The smaller Superdex 200 5/150 column enables more rapid analysis.
7. A relatively stable and monodispersed membrane protein sample will give a single symmetrical peak in the FSEC chromatogram (Fig. 4).

3.4 Isolation of the Plasmid from *S. cerevisiae* and Analysis of Its DNA Sequence

1. Inoculate 5 ml SC medium without uracil with the *S. cerevisiae* transformant. Grow overnight at 30 °C in the shaking incubator.
2. Harvest the cells in a 1.5 ml tube.
3. Suspend the cell pellet into 500 µl sorbitol buffer.
4. Add 20 µl Zymolyase solution (12.5 mg/ml Zymolyase 20T in sorbitol buffer), and mix well.
5. Incubate at 37 °C for 30 min.
6. Centrifuge at 12,000×*g* for 1 min, and discard the supernatant.
7. Suspend the cell pellet in 250 µl buffer P1 from the QIAprep Miniprep Kit.
8. Add 150 µl dry acid-washed glass beads into the cell suspension.
9. Vortex at maximum speed for 1 min.
10. Add buffer P2. From this point, follow the manufacturer's protocol. Elute the plasmid with 30 µl buffer EB (*see Note 13*).
11. Transform *E. coli* (e.g., DH5α or JM109) with 1–3 µl plasmid solution from yeast and grow on an LB agar medium plate containing 100 µg/ml ampicillin, and isolate the plasmid from *E. coli*.
12. Check the DNA sequence.
13. If the expression level is not sufficient in *S. cerevisiae* even after extensive optimization of the culture conditions, we recommend trying another expression host such as *P. pastoris* or insect cells. Amplify the coding region of the GPCR–GFP–His8, digest it with appropriate restriction enzymes (Fig. 2c), and integrate into a *P. pastoris* expression vector (e.g., pPIC9 series) or insect-cell expression vector (e.g., pFastBac series). We demonstrated that the stabilized receptor variants also showed better characteristics than wild-type receptors in these hosts, and overall, improved results may be anticipated (Fig. 4c, d).

4 Notes

1. Swinging bucket rotor enables collecting yeast cells and removing supernatant easier than use of an angle rotor.
2. The glass beads do not need to be acid washed prior to use in the membrane preparation.
3. A fluorescence detector is generally one of the components of an HPLC system. If an FPLC system is used, consult with the manufacturer of the FPLC system about its compatibility with the fluorescence detector. Ideally, the system is placed in a cold room or a chromatography refrigerator.

4. Detergent stock solution. *n*-Dodecyl- α -D-maltopyranoside (DDM) is the most commonly used detergent for GPCRs. Cholesteryl hemisuccinate (CHS) is incorporated into the detergent solution to stabilize the solubilized GPCR. 10 % (w/v) DDM with 2 % (w/v) CHS is generally prepared as the stock solution. The protocol for the preparation of the DDM/CHS stock solution is found at the JCIMPT web site (<http://jcimpt.scripps.edu>).
5. The following modifications have been adopted for the improvement of the expression and stability of GPCRs: (a) A truncation of flexible long N- or C-terminal residues. (b) A point mutation to tryptophan at the 3.41 position in transmembrane helix 3 [9]. The numbering is based on the general indexed position in the Ballesteros–Weinstein system. (c) Deletion of a long third intracellular loop (i3-loop) or replacement of part of the i3-loop by a stable fusion partner, such as T4 lysozyme [10] or a cytochrome b₅₆₂ mutant [11]. A long i3-loop can potentially become a target of degradation or receptor destabilization on the host cell surface. (d) Alanine scanning mutagenesis and the combination of the mutations [12].
6. The greater the number of PCR fragments that are used, the lower the transformation efficiency becomes [13]. In our experience, the transformation efficiency is about 5,000 cfu per 1 μ g plasmid when transformation is performed with four PCR products. This is enough for the subsequent experiments.
7. If there are too many colonies on the plates, then the colonies remain too small to pick up. This causes a large experimental variability. In this case, dilute the remaining cell suspension and spread on a new plate again.
8. Check more than two colonies for each variant. Because of undigested plasmid or PCR error, colonies without expression may appear.
9. Make a stock of the colony. Drop 2–5 μ l of the overnight culture on a SC medium plate without uracil. Allow to dry and then invert the plates. Grow at 30 °C for 2–3 days. This plate can be stored at 4 °C for up to 2 months.
10. If desired, add chemical chaperons such as dimethyl sulfoxide (DMSO) and glycerol, or decrease the cultivation temperature to 20 °C (grow for 40 h). Chemical chaperons and/or optimization of temperature may result in higher expression.
11. The total expression level can be estimated from the GFP fluorescence. For example, using our plate reader, the intensity of 200 μ l of 1 μ M purified yEGFP in a 96-well plate was 7,500 [rfu] at 490 nm excitation and 525 nm emission. Whole-cell fluorescence was measured in a 200 μ l cell suspension by using cultures at the same cell density as the culture. Therefore, the

expression level per one liter culture [mg/l] was calculated from the following: $(\text{GFP counts of whole cell [rfu]}) / (7,500 \times 10^6 [\text{rfu} \times \text{l/mol}]) \times (\text{molecular weight} \times 10^3 [\text{mg/mol}])$.

12. FSEC analysis can be performed without using a fluorescence detector as follows. Collect 0.2 ml fractions of SEC in a 96-well black plate, and read the fluorescence of the fractions using a fluorescence plate reader. The FSEC profiles shown in Fig. 4 were obtained in this way. In this case, more yeast membrane (from at least 50 ml culture) is required because of the lower sensitivity of the plate reader compared with the use of an online detector.
13. The plasmid solution purified from *S. cerevisiae* is unfavorable for DNA sequencing. Therefore, we recommend transforming *E. coli* and amplifying the plasmid for sequencing.

References

1. Fredriksson R, Lagerstrom MC, Lundin LG, Schioth HB (2003) The G-protein coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol Pharmacol* 63:1256–1272
2. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? *Nat Rev Drug Discov* 5:993–996
3. Newstead S, Kim H, von Heijne G et al (2007) High-throughput fluorescent-based optimization of eukaryotic membrane protein overexpression and purification in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* 104:13936–13941
4. Kawate T, Gouaux E (2006) Fluorescence-detection size-exclusion chromatography for precrystallization screening of integral membrane proteins. *Structure* 14:673–681
5. Shiroishi M, Tsujimoto H, Makyio H et al (2012) Platform for the rapid construction and evaluation of GPCRs for crystallography in *Saccharomyces cerevisiae*. *Microb Cell Fact* 11:78
6. Shimamura T, Shiroishi M, Weyand S et al (2011) Structure of the human histamine H(1) receptor complex with doxepin. *Nature* 475:65–70
7. Kota J, Gilstring CF, Ljungdahl PO (2007) Membrane chaperone Shr3 assists in folding amino acid permeases preventing precocious ERAD. *J Cell Biol* 176:617–628
8. Drew D, Newstead S, Sonoda Y et al (2008) GFP-based optimization scheme for the overexpression and purification of eukaryotic membrane proteins in *Saccharomyces cerevisiae*. *Nat Protoc* 3:784–798
9. Roth CB, Hanson MA, Stevens RC (2008) Stabilization of the human beta2-adrenergic receptor TM4-TM3-TM5 helix interface by mutagenesis of Glu122(3.41), a critical residue in GPCR structure. *J Mol Biol* 376:1305–1319
10. Rosenbaum DM, Cherezov V, Hanson MA et al (2007) GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. *Science* 318:1266–1273
11. Chun E, Thompson AA, Liu W et al (2012) Fusion partner toolchest for the stabilization and crystallization of G protein-coupled receptors. *Structure* 20:967–976
12. Serrano-Vega MJ, Magnani F, Shibata Y, Tate CG (2008) Conformational thermostabilization of the beta1-adrenergic receptor in a detergent-resistant form. *Proc Natl Acad Sci U S A* 105:877–882
13. Ito K, Sugawara T, Shiroishi M et al (2008) Advanced method for high-throughput expression of mutated eukaryotic membrane proteins in *Saccharomyces cerevisiae*. *Biochem Biophys Res Commun* 371:841–845

Cell-Free Expression of G-Protein-Coupled Receptors

Erika Orbán, Davide Proverbio, Stefan Haberstock,
Volker Dötsch, and Frank Bernhard

Abstract

Cell-free expression has emerged as a new standard for the production of membrane proteins. The reduction of expression complexity in cell-free systems eliminates central bottlenecks and allows the reliable and efficient synthesis of many different types of membrane proteins. Furthermore, the open accessibility of cell-free reactions enables the co-translational solubilization of cell-free expressed membrane proteins in a large variety of supplied additives. Hydrophobic environments can therefore be adjusted according to the requirements of individual membrane protein targets. We present different approaches for the preparative scale cell-free production of G-protein-coupled receptors using the extracts of *Escherichia coli* cells. We exemplify expression conditions implementing detergents, nanodiscs, or liposomes. The generated protein samples could be directly used for further functional characterization.

Key words G-protein-coupled receptors, Proteomicelles, Protein/nanodisc complexes, Proteoliposomes, Co-translational solubilization

1 Introduction

G-protein-coupled receptors (GPCRs) are, like many other membrane proteins, challenging targets for conventional cell-based expression systems. Cell-free (CF) expression technologies have been established in recent times as new alternative option for the production of membrane proteins [1]. General advantages are the elimination of central bottlenecks in membrane protein production such as toxicity, inefficient translocation, and proteolytic degradation as well as the possibility to modulate expression environments with a large number of additives, e.g., amphiphilic compounds for the solubilization of synthesized membrane proteins. In the continuous exchange cell-free (CECF) expression configuration composed out of two fixed volume compartments containing either reaction mixture (RM) or feeding mixture (FM), the amounts of several mg of protein per 1 ml of RM volume could be obtained [2, 3]. For membrane protein production, three basic

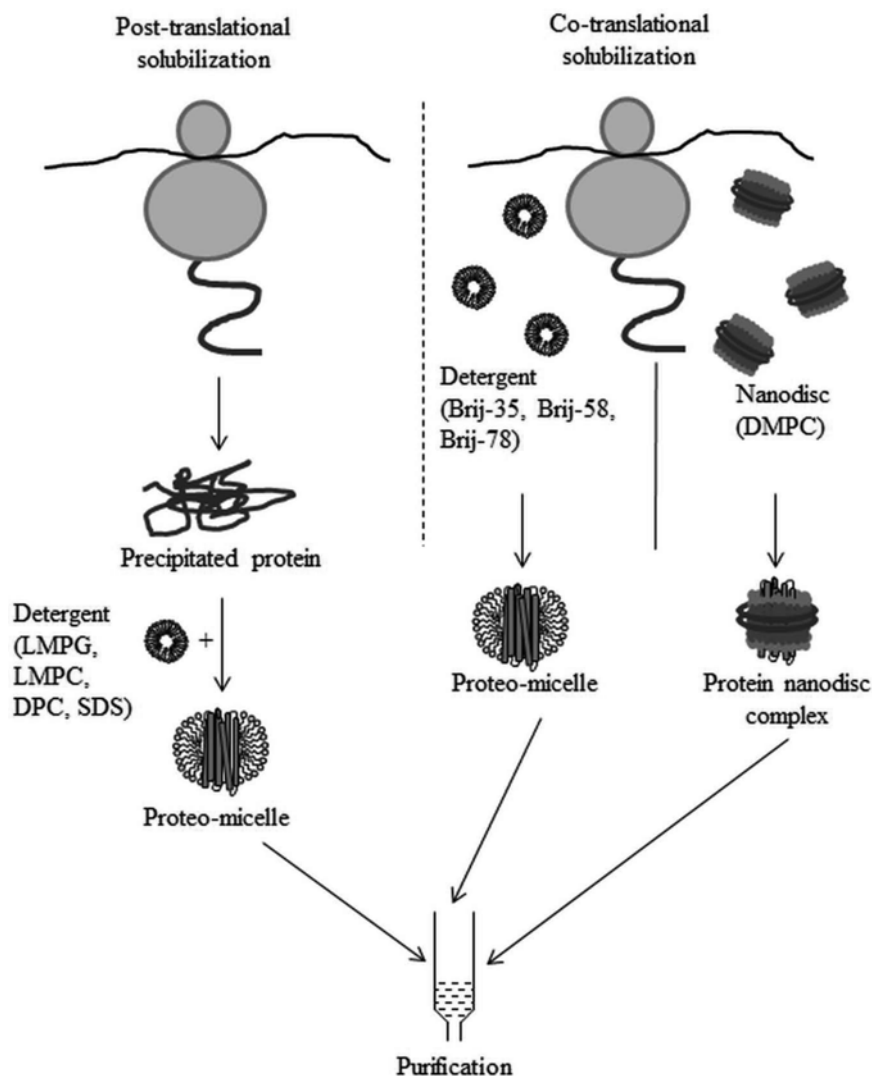


Fig. 1 CF expression approaches for the production of functional GPCRs

CF expression modes can be considered [1, 3]. In reactions without supplied amphiphilic compounds, the freshly synthesized membrane proteins instantly precipitate (P-CF or precipitate-forming mode). The proteins must then be posttranslationally solubilized for further studies (Fig. 1). If micelle- or complex-forming compounds have been added into the reaction mixture, the membrane proteins could be co-translationally solubilized (D-CF or detergent-based mode). In addition, either natural or artificial membranes could be provided (L-CF or lipid-based mode) in order to enable the co-translational insertion of the synthesized membrane proteins (Fig. 1). Which expression mode and what

type of compound or compound mixture is suitable for a particular membrane protein is usually the subject of screening experiments. A variety of different types of membrane proteins have already been successfully produced in CF systems. One major focus has been on members of the GPCR superfamily, and numerous examples have been CF synthesized so far. Most reports cover GPCRs from the dominant class I or rhodopsin-like family such as muscarinic receptors [4], the β 2-adrenergic 2 receptor [5], olfactory receptors [6–9], the histamine H1 receptor [10], endothelin receptors [11, 12], and many others [3, 13]. Even class II receptors have been CF synthesized, and efficient production protocols have been reported [14]. In almost all cases, bacterial CF extracts from *E. coli* have been used for GPCR production, but different expression modes and a diverse variety of solubilization compounds have been implemented (Table 1). As representative examples, we describe CF expression protocols which are efficient for the production of human GPCRs such as the endothelin A and B receptors, the gonadotropin-releasing hormone receptor (GnRHR), and the chemokine receptor CXCR2. The protocols could be used as starting point and guidelines for the CF expression of many other GPCRs as well. However, DNA template design as well as types and concentrations of solubilizing compounds may need to be modified according to individual requirements.

2 Materials

All stock solutions should be prepared with ultrapure water and stored at -20°C if not otherwise stated.

2.1 General Materials

1. Fermenter for bacterial cultures, e.g., 5–10 l volume.
2. French press.
3. Photometer.
4. Thermo shaker for incubation.
5. GFP assay buffer: 20 mM Tris-HCl (pH 7.8) and 150 mM NaCl.
6. Dark microplates (96F Nunclon Delta Black Microwell SI, Nunc).
7. Chromatographic system (e.g., Äkta purifier, GE Healthcare).
8. Q-Sepharose column (GE Healthcare).
9. Immobilized Metal Affinity Chromatography (IMAC) material or column.
10. Ultrasonic water bath.
11. Sonicator.

Table 1
CF expression protocols for the production of GPCRs

GPCR	Size [kDa]	Assay ^a	Mode [mg/ml] ^b	Solubilization compounds [concentration] ^c	Ref.
hETB, hNPY2/5, hMTNRI1A/B, hSS1/2, hV1BR, hHRH1, hV2R, rCRF	39–51	–	P-CF [3]	e.g., LMPG [2 %]	[3]
hCRFR1, mCRFR2βA	47–49	+	P-CF [3], D-CF	LMPG [2 %]/Nvoy or only Nvoy	[14]
hETA/B	49	+	P-CF [3], D-CF [3]	LPPG [1 %], LMPC [1 %], SDS [1 %], Fos16 [1 %], Fos12 [1 %], Brij-35/78 [0.6 %]	[11]
hH1R	56	+	P-CF [1]	DDM [2 %]	[10]
Dopamine D2	50	+	D-CF		[25]
hTAAR-T4L	45	+	D-CF [2]	Brij-35	[26]
hORs, hFPR3, hVN1R1, hVN1R5	~30	+	D-CF [~0.2]	Brij-35 [0.2 %], peptide surfactants [0.5–2.5 mM]	[6]
hOR17-4	36	+	D-CF [2]	e.g., digitonin [0.36 %]	[8]
hMTNRI1B, hNPY4R, rCRF, hV2R	~40	–	D-CF [2]	e.g., Brij-58 [1.5 %], Brij-78 [1 %]	[13]
hCHRM2, hβ2AR, hNTR	~60	+	D-CF [2]	Brij-35 [<1 %], digitonin [<1 %]	[4]
hOR17-210, mOR103-15, hFPR3, hTAAR5	~40	+	D-CF [2]	Peptide surfactants [0.5–2.5 mM]	[7]
hβ2-AR-T4L	63	+	L-CF [?]	apoA-nanodiscs (DMPC)	[5]
hETB, hETA	43–49	+	P-CF [2], D-CF [2], L-CF [2]	e.g., LMPG [1 %], Brij-78 [1 %], MSP1E3-nanodiscs (DMPC/G) [40 μM]	[12]
ADRB2, DRD1, NK1R	45	+	L-CF	Δ49A1-nanodiscs (DMPC)	[27]
OR5	35.5	–	L-CF [0.2–1.8]	DOPC, PC [200 μM]	[9]

^aFunctional assays have been performed

^bReported yields are given in mg of expressed GPCR per ml of reaction mixture

^cConcentrations of compounds in the reaction mixtures are given, if documented

[?]not published

12. Centriprep filter devices, 10 kDa MWCO (Millipore).
13. Thermocycler for polymerase chain reaction (PCR).
14. Mini-extruder (Avanti Polar Lipids).

2.2 Materials for S30 Extract Preparation

1. *E. coli* A19 strain.
2. LB medium: 10 g/l peptone, 5 g/l yeast extract, 10 g/l NaCl.
3. 2× YTPG medium: 10 g/l yeast extract, 16 g/l tryptone, 5 g/l NaCl, 100 mM glucose, 22 mM KH₂PO₄, 40 mM K₂HPO₄.
4. Antifoam (Sigma).
5. 40× S30-A/B buffer: 400 mM Tris-acetate (pH 8.2), 560 mM Mg(OAc)₂, 2.4 M KCl.
6. 1× S30-A buffer (washing buffer): diluted from the 40× S30-A/B stock, supplemented with 6 mM β-mercaptoethanol.
7. 1× S30-B buffer (lysis buffer): diluted from the 40× S30-A/B stock, supplemented with 1 mM DTT and 1 mM phenylmethanesulfonylfluoride (PMSF).
8. 40× S30-C buffer: 400 mM Tris-acetate (pH 8.2), 560 mM Mg(OAc)₂, 2.4 M KOAc.
9. 1× S30-C buffer (dialysis buffer): diluted from the 40× S30-C stock, supplemented with 0.5 mM DTT.
10. 5 M NaCl.

2.3 Materials for T7 RNA Polymerase Preparation

1. *E. coli* BL21 (DE3) Star×pAR1219 [15].
2. LB medium: 10 g/l peptone, 5 g/l yeast extract, 10 g/l NaCl.
3. 1 M isopropyl β-D-1-thiogalactopyranoside (IPTG).
4. 30 % (w/v) streptomycin sulfate.
5. Buffer-T7RNAP-A (equilibration buffer): 30 mM Tris-HCl (pH 8.0), 50 mM NaCl, 1 mM EDTA, 10 mM β-mercaptoethanol, 5 % glycerol.
6. Buffer-T7RNAP-B (dialysis buffer): 10 mM K₂HPO₄-KH₂PO₄ (pH 8.0), 10 mM NaCl, 0.5 mM EDTA, 1 mM DTT, 5 % glycerol.
7. Resuspension buffer: 30 mM Tris-HCl (pH 8.0), 10 mM EDTA, 50 mM NaCl, 5 % glycerol (v/v), and 10 mM β-mercaptoethanol.

2.4 Materials for DNA Template Preparation

1. Specific primers designed for the target DNA.
2. Vent polymerase (New England Biolabs).
3. PCR purification kit (Qiagen).
4. Restriction enzymes and ligase for template preparation.
5. Plasmid DNA purification kit (Macherey-Nagel).
6. Agarose (Rotigrose, Roth).

2.5 Materials for CEFC Expression Reactions

1. 24-well microplates (Greiner).
2. Dialysis tubes, 12–14 kDa MWCO (Spectrum).

3. Reaction containers: analytical scale Mini-CECF reactors or preparative scale Maxi-CECF reactors [3]; D-TubeTM dialyzer, 12–14 kDa MWCO (Merck Biosciences); Slide-A-Lyzer, 10 kDa MWCO (Pierce).
4. Stock solutions required for CECF reactions are listed in Table 2. Chemicals are from Sigma-Aldrich if not otherwise stated.

2.6 Hydrophobic Compounds for Membrane Protein Solubilization

1. Hydrophobic compounds used for the solubilization of G-protein-coupled receptors (GPCRs) are listed in Table 3.
2. Liposome buffer: 20 mM potassium phosphate (pH 7.0), 150 mM NaCl.

2.7 Materials for Nanodisc (ND) Preparation

1. pET-28-MSP1E3D1 vector [16].
2. BL21 (DE3) Star cells.
3. 10 % (w/v) glucose stock solution.
4. PMSF dissolved in ethanol (final concentration should be 1 mM).
5. 10 % (v/v) Triton X-100 stock solution.
6. MSP-A buffer: 40 mM Tris-HCl (pH 8.0), 300 mM NaCl, 1 % (v/v) Triton X-100.
7. MSP-B buffer: 40 mM Tris-HCl (pH 8.9), 300 mM NaCl, 50 mM cholic acid.
8. MSP-C buffer: 40 mM Tris-HCl (pH 8.0), 300 mM NaCl.
9. MSP-D buffer: 40 mM Tris-HCl (pH 8.0), 300 mM NaCl, 50 mM imidazole.
10. MSP-E buffer: 40 mM Tris-HCl (pH 8.0), 300 mM NaCl, 300 mM imidazole.
11. MSP-F (dialysis) buffer: 40 mM Tris-HCl (pH 7.4), 300 mM NaCl, 0.5 mM EDTA.
12. DMPC-cholelate stock solution: 50 mM DMPC, 100 mM sodium cholate.
13. 10 % DPC stock solution (for complete solubilization, ultra-sonic water bath is required).
14. ND-A buffer: 10 mM Tris-HCl (pH 8.0), 100 mM NaCl.
15. Bio-Beads (Bio-Rad).

2.8 Materials for the Purification of CF-Expressed GPCRs

1. Ni-NTA Superflow resin (Qiagen).
2. NPI-10 buffer: 50 mM NaH₂PO₄ (pH 8.0), 300 mM NaCl, 10 mM imidazole.
3. NPI-50 buffer: 50 mM NaH₂PO₄ (pH 8.0), 300 mM NaCl, 50 mM imidazole.
4. NPI-400 buffer: 50 mM NaH₂PO₄ (pH 8.0), 300 mM NaCl, 400 mM imidazole.

Table 2
Reagents for CECF expression reactions

Compound	Stock conc.	Final conc.
Master mix		
RCWMDE amino acid mix	16.7 mM	1 mM
20-amino-acid mix	25 mM	0.5 mM
Acetyl phosphate (Li^+ , K^+), pH 7.0 ^a	1 M	20 mM
Phospho(enol)pyruvic acid (K^+), pH 7.0 ^a	1 M	20 mM
75× NTP mix, pH 7.0 ^b	90 mM ATP	1.2 mM
	60 mM G/C/UTP	0.8 mM
DTT	500 mM	2 mM
Folinic acid (Ca^{2+})	10 mg/ml	0.1 mg/ml
Complete cocktail (Roche Diagnostics)	50×	1×
Hepes/EDTA, pH 8.0 ^a	24×	1×
Mg(OAc) ^b	1 M	11.1 mM ^c
KOAc	4 M	110 mM ^c
PEG 8000	40 %	2 %
NaN_3	10 %	0.05 %
Optional compounds in Master mix ^d		
Option 1: water	Fill up to the final volume	
Option 2: water	Fill up to the final volume	
Option 3: detergent (e.g., Brij-35)	e.g., 15 %	e.g., 0.6 %
Reaction mix (RM)		
Master mix ^c		28.30 %
Pyruvate kinase	10 mg/ml	0.04 mg/ml
t-RNA (<i>E. coli</i>)	40 mg/ml	0.5 mg/ml
T7RNAP	3.2 mg/ml	0.04 mg/ml
RiboLock	40 U/ μl	0.3 U/ μl
DNA template	0.2–0.5 $\mu\text{g}/\mu\text{l}$	2–20 ng/ μl
<i>E. coli</i> S30 extract	1×	0.35×
Optional compounds in RM		
Option 1: nanodisc (DMPC)	500–1,200 μM	80–100 μM
Option 2: liposome (asolectin)	20 mg/ml	4 mg/ml
Option 3: water	Fill up to the final volume	
Feeding mix (FM)		
Master mix ^c		28.30 %
S30-C buffer	1×	0.35×
20-amino-acid mix	25 mM	0.5 mM
Water		

^aAdjusted with KOH

^bAdjusted with NaOH

^cFinal total concentrations are 16 mM Mg^{2+} and 270 mM K^+ as additional amounts of the ions are contributed by other compounds

^dFrom optional compounds, only the low-molecular-weight detergents are present in both compartments, RM and FM. Liposomes or nanodiscs do not pass the membrane and can be added into the RM only. Examples for the three alternative options of the addition of nanodiscs, liposomes, or detergents are given

^eMaster mix should be added to the RM and FM. Final concentration in this case means that Master mix should be 28.3 % of the final RM or FM volume

Table 3
Hydrophobic compounds for the solubilized of CF-expressed GPCRs

Compound	Stock concentration	Working range	Recommended use
Detergents			
Brij-35	15 % (w/v) ^a	0.6 %	D-CF
Brij-58	15 % (w/v) ^a	1.5 %	D-CF
Brij-78	15 % (w/v) ^a	1.0 %	D-CF
DPC	10 % (w/v) ^a	0.2 %	D-CF, P-CF
DDM	10 % (w/v) ^a	0.1 %	D-CF, P-CF
LMPC	10 % (w/v) ^a	2 %	P-CF
LMPG	10 % (w/v) ^a	2 %	P-CF
SDS	10 % (w/v) ^a	2 %	P-CF
Membranes			
Liposome (asolectin)	20 mg/ml	4 mg/ml	L-CF
Nanodisc (DMPC)	500–1,200 μ M	80–100 μ M	L-CF

^aDissolved in water

3 Methods

3.1 S30 Extract Preparation

The *E. coli* strain A19 is one of the most frequently used for the preparation of S30 extracts since it has low RNase activity. Starting with a 10 l fermentation in 2 \times YT medium yields approximately 60 ml of S30 extract. Strain A19 does not carry selection markers and risk of contamination is therefore increased. Check the purity of the pre-culture and the final fermenter culture by plating aliquots on LB agar plates. The main steps of S30 extract preparation are:

1. Plate *E. coli* A19 strain on agar plate and incubate for approximately 7–9 h at 37 °C (*see* **Note 1**).
2. Inoculate 120 ml LB media and shake overnight at 37 °C (pre-culture).
3. Inoculate a fermenter containing 10 l 2 \times YPTG medium with 100 ml of the pre-culture.
4. Incubate at 37 °C with continuous stirring (500–700 rpm) and vigorous aeration until mid log phase (OD600 approximately 3.5–4.5; *see* **Note 2**).
5. Cool down from 37 to 18 °C in a maximum time of 30–40 min.
6. Harvest and centrifuge the cells at 6,800 $\times g$ for 15 min.
7. Wash the pellet with S30-A buffer, centrifuge at 8,000 $\times g$ for 10 min.
8. Repeat this step two more times.
9. Weigh the pellet and suspend it in 110 % S30-B buffer.

10. Disrupt cells with French press slowly, and at high pressure, the solution should become grayish. Centrifuge the solution at $30,000 \times g$ for 30 min.
11. Transfer the supernatant into a fresh tube. Centrifuge the supernatant one more time at $30,000 \times g$ for 30 min.
12. Transfer the supernatant to a fresh tube and adjust to a final concentration of 400 mM NaCl. Remove endogenous mRNA as well as undesired proteins using high salt concentration (400 mM NaCl) and heating to 42 °C for 45 min.
13. Dialyze the turbid solution overnight against S30-C buffer using a 12–14 kDa cutoff membrane.
14. Pour the turbid solution into a centrifuge tube and centrifuge at $30,000 \times g$ for 30 min.
15. Transfer the supernatant into a fresh tube and repeat centrifugation once.
16. Harvest the supernatant, mix quickly, and aliquot into suitable volumes. The final total protein concentration of the extract should be at least 30 mg/ml. Shock-freeze in liquid nitrogen and store at –80 °C. Frozen extract is stable for many months. Aliquots can be thawed on ice and left on ice for few hours during the setup of reactions. Unused extract can be refrozen, but efficiencies might become lower.
17. Check the efficiency of each S30 extract batch with the expression of GFP or any other suitable protein standard and determine the basic Mg^{2+} ion optimum. The Mg^{2+} ion optimum is usually within the range of 12 mM up to 24 mM.

3.2 T7RNAP Preparation

1. T7RNAP is produced from the *E. coli* strain BL21 (DE3) Star (pAR1219) by conventional cultivation in Erlenmeyer flasks with LB medium. The steps of T7RNAP preparation are listed below:
2. Prepare pre-culture from BL21 (DE3) Star × pAR1219, carrying the gene for the T7RNA polymerase, and shake it overnight at 37 °C.
3. Inoculate 1 l LB medium 1:100 with the pre-culture.
4. Incubate the culture at 37 °C on a shaker until $OD_{600}=0.6-0.8$.
5. Induce T7RNAP expression with 1 mM IPTG.
6. Incubate for further 5 h at 37 °C.
7. Harvest the cells by centrifugation at $4,500 \times g$ for 15 min.
8. Resuspend the pellet in 30 ml resuspension buffer.
9. Disrupt the cells with French press, centrifuge at $20,000 \times g$ for 30 min, and transfer the supernatant into a fresh tube.
10. Precipitate nucleic acids in the supernatant with 4 % streptomycin sulfate for 5 min on ice.

11. Centrifuge the lysate at $20,000 \times g$ for 30 min.
12. Load the supernatant onto a 40 ml Q-Sepharose column equilibrated with 2 column volumes (CV) of equilibration buffer.
13. Wash the column with equilibration buffer at a flow rate of 4 ml/min.
14. Elute bound proteins with a gradient from 50 to 500 mM NaCl at a flow rate of approximately 3 ml/min.
15. Check for a prominent band at approximately 90 kDa by SDS-PAGE analysis and Coomassie staining. Pool the fractions with highest T7RNAP content (*see Note 3*). On average, approximately 20,000–40,000 T7RNAP units can be isolated out of 1 l culture.
16. Dialyze pooled fractions against dialysis buffer.
17. Concentrate T7RNAP to 1–4 mg/ml by ultrafiltration (*see Note 4*).
18. Aliquot T7RNAP solution in dialysis buffer, add glycerol to a final concentration of 50 % (v/v), and store at -80°C . Stored aliquots are stable for many months. Working aliquots could be stored at -20°C .

3.3 DNA Template Design and Preparation

3.3.1 Basic DNA Template Design

The reading frame to be expressed has to be under control of the T7 promoter and T7 terminator. For efficient expression, high quality and purity of the template DNA are critical. The optimal template concentration should be determined for each new target with an initial concentration screen in the range of 0.1–20 ng/ μl RM. As a template of expression reactions, linear DNA or plasmid DNA containing the gene of interest can be used. Commonly used plasmid vectors are the pET (Merck Biosciences) or pIVEX (Roche Diagnostic) series:

1. Plasmid templates should be prepared with commercial standard kits such as “Midi” or “Maxi” DNA purification kits according to the manufacturers’ instructions (*see Note 5*). At the end of the procedure, the DNA should be dissolved in pure Milli-Q water. Optimal final template concentrations are in between 0.2 and 0.5 mg/ml.
2. Linear target DNA can be prepared by PCR. In this case, the PCR product should be purified using a commercial PCR purification kit. This option can be used only if the target gene is already present in a suitable vector under control of T7 promoter elements. If the target gene is under control of a different promoter, the appropriate T7 regulatory elements must be attached by a multistep overlap PCR strategy before further utilization [17].

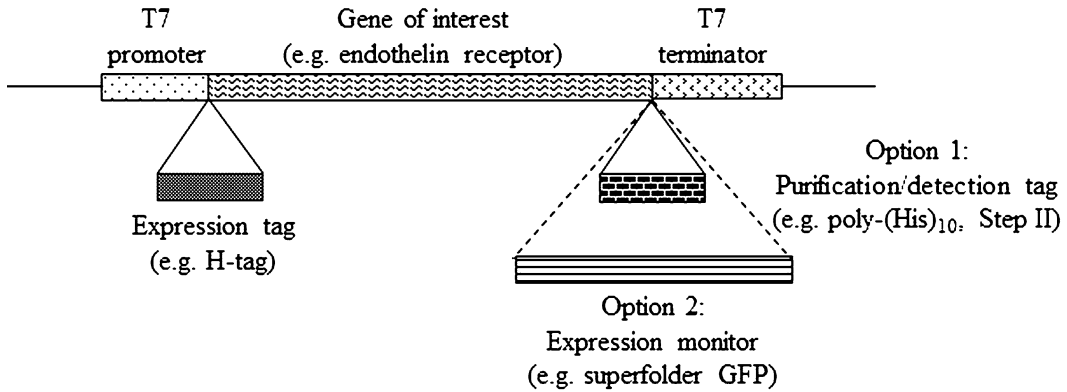


Fig. 2 DNA template design for the CF expression of GPCRs. Different optional modifications (i.e., expression tags, purification tags, or expression monitors) are indicated

3.3.2 Optional Template Modifications

Expression tags can be added to the basic template design either to try and improve expression levels or simply to enable rapid monitoring or detection (Fig. 2). Expression tags are small additional N-terminal sequences comprising up to six codons [17]. Their optimized nucleotide sequences minimize secondary structure formation of the mRNA resulting into improved initiation of translation. Expression tags are the primary choice if low expression persists even after initial reagent optimization (e.g., Mg^{2+} ions). We recommend evaluating two tags, the AT-tag (AAA TAT TAT AAA TAT TAT) and/or the H-tag (AAA CCA TAC GAT GGT CCA). In our hands, the addition of one of these tags boosted expression efficiencies up to sufficient yields in the vast majority of cases. Expression tags might be cleaved off by the implementation of protease cleavage sites, or in some cases, they could be reduced down to one to three codons while still retaining high expression yields [17]. Alternatively, the optimization of the first natural 5-prime codons of the mRNA could result into similar expression-enhancing effects. If expression yields remain low despite the addition of expression tags, *E. coli* codon-optimized synthetic genes might be considered.

The detection and purification of CF-expressed proteins could be supported by the addition of purification/detection tags (e.g., poly-(His)₁₀ tag or Strep II tag) to the C-terminal end of the target protein (*see* Fig. 2).

Fusions with monitoring systems such as derivatives of the green fluorescent protein (GFP) might be useful for expression protocol optimization [18]. The GFP partner should be attached to the C-terminus of the target protein as in some cases the functional folding of the C-terminal GFP moiety may correlate with the folding of the N-terminal target protein [19]. We recommend fusions with superfolder GFP [$\lambda Ex = 484/\lambda Em = 510$] as it has a

significantly higher folding efficiency in the presence of many detergents compared with wild-type GFP or with shifted GFP derivatives [18]. Monitoring systems are highly valuable in compound screens for the co-translational solubilization of membrane proteins. They are not useful in posttranslational solubilization screens as the superfolder GFP moiety usually unfolds upon precipitation of the N-terminally attached membrane protein fusion partner. The expression of superfolder GFP or of membrane protein fusions with superfolder GFP can be quantified using the following protocol:

1. Express superfolder GFP or membrane protein–superfolder GFP fusion.
2. Add 3 μ l of sample (e.g., supernatant of the RM after reaction) into 297 μ l GFP assay buffer in a 96-well dark microplate and shake for a short time (*see Note 6*). In case of low expression efficiency, higher volumes of sample can be taken.
3. Measure fluorescence at the appropriate wavelength.
4. For quantification, use a calibration curve with purified superfolder GFP.

3.4 Continuous Exchange Cell-Free (CECF) Expression

A semipermeable membrane (Spectrum, 12–14 kDa MWCO) separates the RM from the FM in the CECF expression configuration. CECF expression reactions can be performed in analytical or in preparative scales. Analytical scale reactions in 50–100 μ l volumes are required for protocol development or for screening experiments and can be operated in reusable Mini-CECF reactors [3] (*see Note 7*) or in commercially available D-Tube (Novagen) containers (*see Note 8*). The containers hold the RM and need to be placed in a suitable compartment holding the appropriate volume of FM. We recommend 24-well microplates for the Mini-CECF reactors (*see Note 9*) and 2 ml Eppendorf tubes for small analytical scale D-Tube dialyzer. Established protocols can then be scaled up to several ml of RM in preparative scale reactions. Those reactions can be carried out in Maxi-CECF reactors having commercial Slide-A-Lyzer devices (Pierce) as RM container (*see Note 10*), in larger D-Tube containers, or simply in dialysis tubes placed into 15–50 ml Falcon tubes.

3.4.1 Basic CECF Protocol

A basic flow chart for CECF reactions is illustrated in Fig. 3. The most economical RM:FM ratios are in between 1:14 and 1:20 (*see Note 11*):

1. Calculate the individual compound volumes according to the planned number of reactions and determine the total volumes of RM and FM.
2. Calculate each reagent volume required for the Master mix (Table 2), combine in a tube, and vortex briefly.

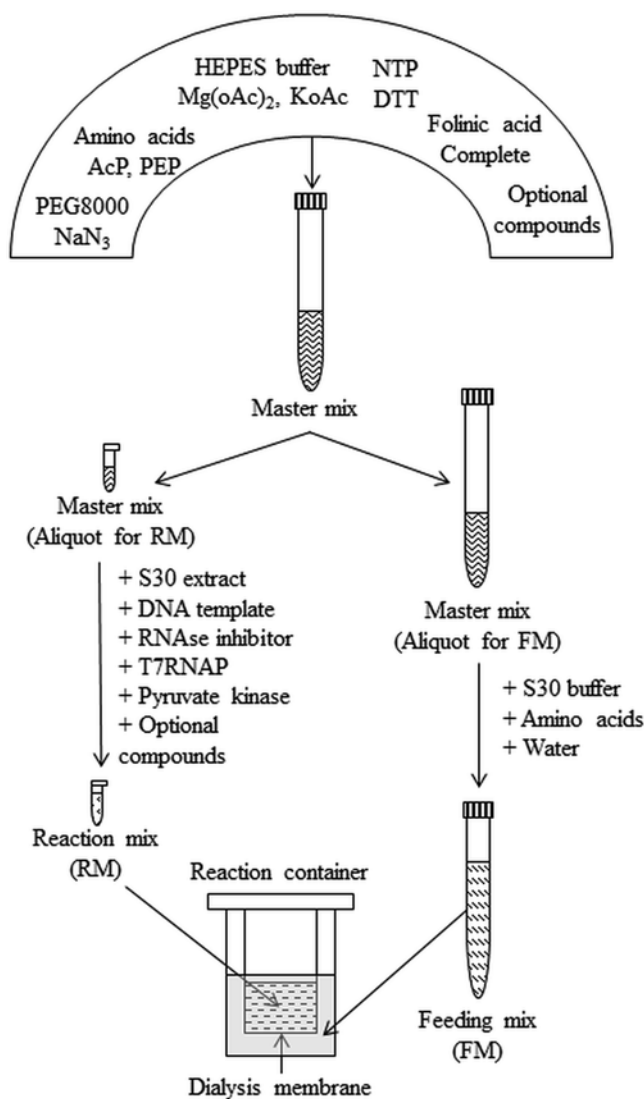


Fig. 3 Flow chart for the preparation of the CECF reaction mixture and feeding mixture. For more details and applied concentrations, *see* Table 2

3. Split the Master mix into the appropriate aliquots for RM and FM.
4. Complete RM and FM with the reagents listed in Table 2 and finally fill up with Milli-Q water (*see* Note 12).
5. Transfer the RM and FM aliquots into the selected reaction containers.
6. Incubate CECF reactions overnight at 30 °C with continuous shaking in order to ensure efficient substance exchange between RM and FM through the membrane (*see* Note 13).

7. Optional: for compound screens comprising a series of analytical CECF reactions, Master mixes can be preprepared with the lowest screening concentration of the compound (*see* **Note 14**).

3.4.2 GPCR Expression Yield Optimization

Efficient CECF expression protocols are first established in the P-CF mode. Most important parameters are (a) DNA template design, (b) DNA template concentration, and (c) Mg^{2+} ion concentration. First, the optimal design of the DNA template has to be determined, and it has to be decided whether N-terminal expression tags and/or C-terminal purification/detection tags are added to the protein (*see* **Note 15**). Starting material could be in principle any available construct under control of T7 regulatory elements.

1. Add a new DNA template at a final concentration of 15 ng/ μl into the RM.
2. CECF reactions should be performed in duplicates and at three different final Mg^{2+} concentrations, usually 14, 16, and 18 mM.
3. After incubation, spin down the RM in a tabletop centrifuge at $18,000\times g$ for 10 min at room temperature.
4. As a first estimate, the GPCR expression correlates with the formation of a pellet. Wash the pellet with 1 ml of S30 buffer and centrifuge again. If no or only small pellets are detectable, proceed with the optimization of the DNA template design (*see* Subheading 3.3.2).
5. Suspend washed pellets in appropriate volumes of S30-C buffer; maximal volumes are the initial RM volume. Analyze appropriate aliquots of the suspension by SDS-PAGE and identify expressed target protein by Western blot, if applicable (*see* **Note 16**).
6. Select Mg^{2+} ion concentration giving the highest yield and proceed with DNA template concentration optimization. Perform CECF reactions with final template concentrations of 0.15, 0.3, 0.6, 1.2, 2.4, 5, 10, and 20 ng/ μl . Evaluate the resulting precipitates and proceed with best template concentration.

3.4.3 GPCR Solubilization in Detergents

Detergents for GPCR solubilization can be used co- or posttranslationally. The type of the detergent and the applied concentrations depend on the solubilization strategy.

For co-translational solubilization (D-CF mode), only mild detergents (e.g., Brij-35, Brij-58, Brij-78) are used, while P-CF-produced protein precipitates have to be solubilized using harsher detergents (e.g., SDS, LMPG, LMPC, DPC, DDM). The quality and stability of the solubilized protein often differ considerably with the various detergents. The most commonly employed detergents and their recommended concentration ranges are listed in Table 3.

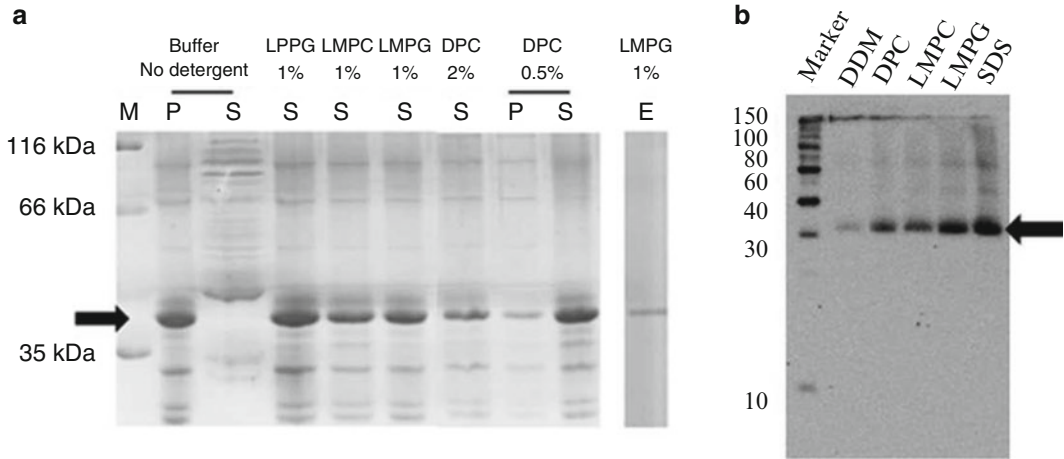


Fig. 4 Examples of posttranslational solubilization screens of P-CF-expressed GPCRs. Precipitates were solubilized in the indicated detergents and analyzed by 16 % SDS-PAGE. *Arrows* indicate the expressed GPCRs. **(a)** Coomassie staining of 1 % detergent-solubilized samples of the human endothelin B receptor (S, soluble fraction; P, pellet). E: Elution fraction of LMPG-solubilized endothelin receptor after Ni-NTA purification. **(b)** Western blot after solubilization of the human GnRHR receptor in 2 % detergent, only the soluble fractions are shown. Marker proteins are indicated in kDa

Most efficient in the solubilization of P-CF-generated precipitates are SDS > LMPG = LMPC > DPC > DDM.

Besides detergents, other hydrophobic compounds such as peptide surfactants [7], nonionic amphipols [20], or fluorinated surfactants [21] might be useful as well for the solubilization of GPCRs, but will not be discussed in this chapter.

Posttranslational Solubilization of GPCRs from P-CF-Generated Precipitates. The detergent profile for the re-solubilization of P-CF-produced precipitates depends on the individual membrane protein targets and can be determined by an initial detergent screen (Fig. 4):

1. Set up a 1 ml CECF reaction in the P-CF expression mode (see Subheading 3.4.1).
2. After the reaction, centrifuge the RM at $18,000 \times g$ for 10 min at room temperature.
3. Resuspend the pellet in 20 mM Tris-HCl (pH 7.0), 10 mM DTT, 100 mM NaCl.
4. Divide the suspension into a suitable number of aliquots and transfer each aliquot into a fresh tube. Centrifuge again at $18,000 \times g$ for 10 min at room temperature.
5. Discard the supernatants and suspend the pellets in the selected detergent solutions (e.g., LMPG, LMPC, SDS, DDM, or DPC at final detergent concentrations of 2 % (w/v)). The volume of the suspension should not exceed the original RM volume,

e.g., if the precipitate of a 1 ml reaction is divided into ten aliquots, the suspension volume of each aliquot should be maximal (100 μ l).

6. Incubate the protein–detergent mixture at 30 °C with slight agitation for 2 h (*see* **Note 17**).
7. After incubation, centrifuge the solution at 18,000 $\times g$ for 10 min at room temperature. Transfer the supernatants into fresh tubes and analyze aliquots of supernatant and residual pellets by SDS-PAGE (for examples, *see* Fig. 4).

Co-translational Solubilization of D-CF-Expressed GPCRs. For the co-translational solubilization of GPCRs in the D-CF mode, mild detergents such as Brij-35, Brij-58, or Brij-78 are used. Detergent stocks are 15 % (w/v) in water. Recommended initial final detergent concentrations are shown in Table 3 and may be subject of later refinement. The detergent profile for the co-translational solubilization of individual membrane protein targets is determined by a detergent screen. The implementation of a membrane protein–superfolder GFP fusion can significantly facilitate the screening process by measuring the resulting fluorescence in the RM supernatant. If suitable detergents have once been identified, subsequent expression reactions could then be performed with the non-modified GPCR:

1. Set up a 1 ml CECF reaction (*see* Subheading 3.4.1) and replace a suitable part of the water volume in RM and FM by the selected detergent stock.
2. After the reaction, centrifuge the RM at 18,000 $\times g$ for 10 min at room temperature.
3. Transfer the supernatants into fresh tubes and analyze appropriate aliquots by SDS-PAGE and/or Western blotting (for examples, *see* Fig. 5).
4. In addition or alternatively, if superfolder GFP fusions have been used, measure the fluorescence of the supernatant (*see* Subheading 3.3.2).
5. If detergents suitable for the solubilization have been identified, further systematic concentration screens may be performed in order to define the required amount of detergent necessary for complete solubilization.

3.4.4 GPCR Expression in the Presence of Liposomes

Lipids are almost generally highly tolerated by CECF reactions, and membranes could be added to the RM (reaction mix) as liposomes or nanodiscs (ND). It has to be considered that the solubilization of an expressed membrane protein does not necessarily indicate its proper reconstitution into the provided membranes. Simple attachment or partial insertions might occur as well and result into nonfunctional proteins. The expression of

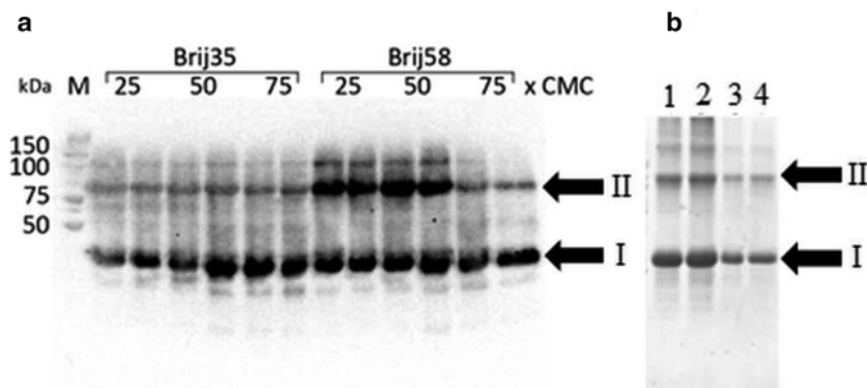


Fig. 5 Co-translational solubilization screen and purification of the human CXCR2 receptor. Samples were analyzed by 16 % SDS-PAGE and stained with Coomassie blue. *Arrows* indicate the expressed GPCR (I, monomer; II, dimer). **(a)** CXCR2 was expressed in the presence of 25 \times , 50 \times , and 75 \times CMC Brij-35 and Brij-58. **(b)** CXCR2 was expressed in D-CF mode with Brij-35. Sample was loaded on Ni-NTA, washed with 70 mM imidazole, and eluted with 350 mM imidazole. Numbers represent different elution fractions

fusions with GFP can also be very helpful in this context. The selected lipids or lipid mixtures can have a significant impact on the solubilization/insertion efficiency of the expressed membrane protein. In our experience so far, soybean asolectin mixture or DMPC has been most efficient for the co-translational solubilization of GPCRs. However, lipid screens might be necessary for particular targets.

Liposomes are prepared *in vitro* and added into the RM of CECF reactions. Despite sizing to 0.2 μ m by extrusion, the liposomes almost quantitatively precipitate during the CECF reaction, presumably by fusion into larger vesicles. If membrane protein fusions with GFP have been expressed, fluorescence can therefore be measured in the pellet. Almost all of the expressed membrane protein (i.e., associated with the liposomes or precipitated) is present in the pellet:

1. Solubilize appropriate amounts of lipid (e.g., asolectin) in 100 % chloroform.
2. Evaporate the chloroform under vacuum overnight until lipids remain as a thin film.
3. Resuspend the lipid film in liposome buffer to a final concentration of 20 mg/ml.
4. Extrude the liposomes to a size of 0.2 μ m using a Mini-extruder (*see Note 18*).
5. Add extruded liposomes directly to the RM to a final concentration of 4 mg/ml.

6. After incubation, separate the supernatant from precipitates by centrifugation of the RM at $18,000\times g$ for 10 min at room temperature.
7. The pellets containing protein precipitates, empty liposomes, and potentially proteoliposomes may be washed with 4 M urea or separated by sucrose density gradient centrifugation before further analysis.

3.4.5 GPCR Expression in the Presence of Nanodiscs (NDs)

Nanodiscs (NDs) are relatively stable natural membrane mimetics composed out of derivatives of the membrane scaffold protein (MSP) and lipids (e.g., DMPC) in a defined ratio. The size of the NDs is determined by the MSP type selected for the preparation. For the co-translational formation of GPCR/ND complexes, we recommend using the MSP1E3D1 derivative which has an average diameter of 12 nm and DMPC as the initial lipid selection in a ratio of 1:115 [22].

NDs need to be preformed and are added into the RM before starting the expression. CF expression in the presence of NDs is an excellent way of screening for the effect of membrane lipid compositions on the expression of a membrane protein [12]. If NDs filled with other lipids and/or composed from different MSP derivatives are desired, the ND preparation protocol must be modified accordingly [23]. The NDs supplemented into CECF reactions are more stable compared with liposomes and remain soluble. Generated GPCR/ND complexes can therefore be purified from the supernatant of reactions by, e.g., affinity chromatography:

Expression of the MSP1E3D1 protein:

1. Transform BL21 (DE3) Star cells with the pET28b-MSP1E3D1 construct [16].
2. Inoculate a 2 l flask containing 750 ml LB medium supplemented with 30 µg/ml kanamycin with the transformed cells and incubate the culture overnight at 37 °C on a shaker (pre-culture).
3. Inoculate 8×2 l flasks each containing 600 ml LB medium supplemented with 0.5 % (w/v) glucose and 3 % (w/v) kanamycin with 50 ml of the pre-culture. Incubate the flasks at 37 °C on a shaker until OD₂₈₀ = 1.0.
4. Induce expression with 1 mM IPTG and continue incubation for further 1 h at 37 °C.
5. Reduce temperature to 28 °C and incubate the culture for further 4 h.
6. Harvest cells by centrifugation at $6,000\times g$ for 10 min at 4 °C.
7. Weigh the pellet and store at -20 °C until purification.

Purification of the MSP1E3D1 protein:

1. Prepare 50 ml MSP-C buffer supplemented with 1 mM PMSF and one tablet Complete protease inhibitor (*see Note 19*).
2. Resuspend the pellet resulting from 4×600 ml fermentation in MSP-C buffer. The final volume of the suspension should be 45 ml.
3. Add Triton X-100 to the suspension in a final concentration of 1 % (v/v). For example, use 5 ml 10 % stock solution for the 45 ml suspension.
4. Disrupt cells by ultra-sonication. Sonicate the suspension for 3×60 s and then 3×45 s with 60 s cooling periods in between the sonication cycles.
5. Centrifuge the suspension at $30,000 \times g$ for 20 min at 4 °C.
6. Filter the supernatant through a 0.45 μ m syringe filter.
7. Equilibrate IMAC column with 5 column volumes (CV) of Milli-Q water and 5 CV MSP-A buffer.
8. Load the filtered supernatant on the IMAC column.
9. Wash the column subsequently with 5 CV MSP-A, MSP-B, MSP-C, and MSP-D buffer.
10. Elute MSP1E3D1 with 5 CV MSP-E buffer and collect fractions.
11. Monitor the OD280 absorbance of the fractions and combine all MSP-containing fractions into one tube.
12. Adjust the sample to a final concentration of 10 % glycerol.
13. Dialyze the sample overnight at 4 °C against 5 l MSP-F buffer with one buffer exchange after 2 h.
14. Centrifuge the dialyzed sample at $18,000 \times g$ for 30 min at 4 °C.
15. Transfer the supernatant to a new tube, determine the total protein concentration, and store at -20 °C until further investigation (usually the protein concentration is 80–100 μ M).

Nanodisc assembly:

1. Mix purified MSP1E3D1, DMPC–cholate stock, DPC, and ND-A buffer in a tube. The final concentration of DPC is 0.1 % (w/v), and the MSP1E3D1:DMPC ratio is 1:115 [23].
2. Incubate the mixture at room temperature for 1 h.
3. Add 0.5 g Bio-Beads (equilibrated with ND-A buffer) per ml solution in order to remove the detergent and incubate for further 4 h on a shaker at room temperature. Alternatively, dialysis against ND-A buffer can be used for the removal of detergent.

4. Centrifuge the solution at $18,000\times g$ for 30 min in order to completely remove Bio-Beads from the solution.
5. Fill the supernatant into a Centriprep concentrating unit (MWCO 10 kDa) equilibrated with ND-A buffer and centrifuge at $2,000\times g$ for 20 min several times until the final MSP1E3D1 concentration is approximately 2.4 mM, corresponding to 1.2 mM ND concentration.
6. Aliquot and freeze ND samples in liquid nitrogen and store the aliquots at $-80\text{ }^{\circ}\text{C}$ until further use. The homogeneity of ND samples could be checked by size-exclusion chromatography using Superdex 200 3.2/30 column.

CECF expression in the presence of NDs:

1. The optimal final concentration of NDs in the reaction depends on (a) the expression efficiency of the GPCR and (b) on the efficiency of the GPCR to associate with the NDs. It is therefore recommended to perform an initial ND concentration screen.
2. The steps of the screening are the following:
3. Prepare a ND stock with DMPC as it is described in Subheading 3.4.5, **step 3** (e.g., 500 μM stock).
4. Prepare Master mix for five duplicate cell-free reactions according to Table 2 (e.g., for 5 concentrations: 0, 20, 40, 80, 100 μM).
5. Split Master mix into RM and FM, and complete FM as it is described in Table 2.
6. Supplement RM according to Table 2 and divide into appropriate aliquots.
7. Complete the RMs with corresponding volumes of ND stock and water.
8. Assemble the reaction chamber and incubate the reactions overnight at $30\text{ }^{\circ}\text{C}$ with continuous shaking.
9. Centrifuge RM at $18,000\times g$ for 10 min and analyze the supernatant and pellet by gel electrophoresis and/or Western blot. The pellet fraction should be decreased with increasing ND concentrations. If no or only partial solubilization of the GPCR could be achieved even with the highest ND concentration, NDs assembled with other lipids such as DMPG may be analyzed. If GPCR–GFP fusions have been used, the fluorescence in the supernatant of the reaction can be used for evaluation.

3.5 Purification of CF-Expressed GPCRs

Several alternative strategies are possible, and purification usually takes advantage of affinity purification tags such as poly(His) $_n$ or Strep II tags. For detergent-solubilized GPCRs, both strategies could be used and all buffers have to be supplemented with the

corresponding detergent. GPCR/liposome complexes may be purified by sucrose density gradient centrifugation. If NDs contain a terminal poly(His)_n tag, GPCRs modified with a C-terminal Strep II tag may be used. Purification by taking advantage of the Strep II tag would therefore enrich for GPCR/ND complexes, while samples purified by using the poly(His) tag will contain the GPCR/ND complexes as well as potentially residual empty NDs. We only exemplify the purification of poly(His)-tagged samples using Ni-chelate affinity chromatography:

1. Equilibrate appropriate amounts of Ni-NTA Superflow resin with 5 CV NPI-10 buffer (*see Note 20*).
2. Dilute the sample (if necessary) with NPI-10 buffer and mix it with the equilibrated Ni-NTA resin.
3. Incubate the sample–Ni-NTA mix for 2 h at appropriate temperature with continuous mixing.
4. Wash the resin with 3–5 CV NPI-50 buffer and collect fractions.
5. Elute the protein with NPI-400 buffer (*see Note 21*).
6. The protein concentration and purity of the elution fractions are determined by SDS-PAGE and/or Western blot.

3.6 Current Challenges and Perspectives

The presented protocols were established with only a small number of human GPCRs, for example, the endothelin A and B receptors, the GnRHR, and the CXCR2 receptor, but have been successfully applied to a number of other GPCRs and also other types of membrane proteins [1]. Table 1 compiles reported CF expression conditions for a variety of GPCRs and indicates variations in expression mode and solubilization compounds. A current challenge is to identify the parameters responsible for the integration efficiency of CF-expressed GPCRs into supplied membranes. The implementation of redox systems for the improved formation of essential disulfide bridges, screening of additives for higher long-term stability of GPCR samples, and evaluation of alternative extract background for expression are further subjects of the current research.

Cell-free expression is a rapidly growing technology, and new modifications are continuously emerging. In addition to the hydrophobic environments described in this chapter, new artificial compounds such as peptide surfactants or amphipols have been reported for the solubilization of GPCRs into functional conformations [6, 24]. In addition, specific combinations of lipids with detergents or other compounds might represent perfect media for a particular GPCR species. CF expression is an excellent tool in order to rapidly obtain mg amounts of GPCR protein, and basic expression protocols are usually established within a month. However, it should be noted that the determination of optimal and

target-specific solubilization conditions requires additional time investments and could require a number of screens for suitable detergents, lipids, or mixtures thereof. The availability of in vitro assays for the fast quality control of samples is therefore often indispensable.

4 Notes

1. A freshly grown culture is necessary for inoculation. Do not use older cultures or cultures stored in the cold room for inoculation.
2. The time of harvesting is most important in S30 extract preparation. The indicated OD600 values are examples only. Make sure that the culture is in the mid log phase of growth at your fermentation conditions. Pilot experiments in order to define the OD600 value for mid log growth phase are highly recommended.
3. Significant impurities will still be present in the elution fraction. T7RNAP may smear over several elution fractions. Combine only peak fractions.
4. The total protein concentration should finally be 3–4 mg/ml. T7RNAP may start to precipitate at higher concentrations.
5. “Mini” kit preparations are not suitable as a CECF DNA template due to the low quality of the purified DNA.
6. The dilution of the sample should be adjusted in order to stay within the range of the calibration curve.
7. For the assembly of the Mini-CECF reactor, the Teflon ring is placed on a sheet of parafilm and then a suitable piece of dialysis membrane (2 × 2 cm) is placed on top of the Teflon ring. Finally, the container is pushed through the Teflon ring which tightly fixes the dialysis membrane between the ring and container. The Mini-CECF reactor is filled from the top by touching carefully the membrane with the pipette tip and releasing the RM. For harvesting the RM after incubation, the membrane is perforated from the bottom with a pipette tip and the RM is removed. After the reaction, the Mini-CECF reactor is disassembled and the membrane is disposed. The container and the Teflon ring are cleaned by extensively washing with Milli-Q water. Prior to next usage, the container and Teflon ring should be dried completely.
8. D-Tube dialyzer may be reused few times after extensive washing with water and storage in water with 0.1 % NaN₃. The water must be removed completely before filling with the RM. We recommend reuse only for the same protein target.

9. Microplates with the Mini-CECF reactors should be sealed with parafilm to prevent evaporation during incubation.
10. The Slide-A-Lyzer is filled with a syringe at one of the preformed openings. This opening should be placed upwards if Maxi-CECF reactors are used. It must be sealed if other FM containers are used. Slide-A-Lyzers can be reused a few times after extensive washing with water and storage in water with 0.1 % NaN_3 . We recommend reuse only for the same protein target.
11. Expression efficiency is neither linear with the RM volume nor with the RM:FM ratio. The indicated volumes and ratios are good economical compromises but may be modified if desired.
12. The FM can be vortexed briefly; the RM should only be mixed by inverting or by pipetting up and down.
13. Shaking water baths or thermo-controlled cabinets with shaking plates at approx. 150–200 rpm may be used.
14. The volume of screening compounds needs initially to be subtracted from the water volume of the RM and FM mixtures.
15. The modification of the target protein with tags may not be desired. Nevertheless, monitoring tags such as GFP may still be necessary for establishing an efficient expression protocol. The production of the target could then be performed with templates encoding for the tag-less derivative.
16. In many precipitates, a considerable amount of coprecipitated proteins from the extract is visible. This coprecipitation is probably due to unspecific hydrophobic interactions.
17. The incubation time and temperatures as well as the solubilization volume are protein-dependent parameters.
18. Extruded liposomes could be stored at 4 °C up to 1 week or at –20 °C, but in this case, they should be extruded again before use.
19. PMSF stock solution should be prepared freshly and solubilized in ethanol.
20. The amount of Ni-NTA Superflow resin depends on the amount of the protein wanted to be purified. If analytical scale reaction is set up and the amount of expressed protein is low, the purification might be carried out using spin-off columns. If more protein wanted to be purified, gravity-flow column should be prepared from the Ni-NTA Superflow resin.
21. Elution from the Ni-NTA column might be carried out with increased imidazole concentration or with a pH shift to pH 4.5. In both cases, a dialysis step might be necessary before further experiments depending on the sensitivity of the subsequent assay.

Acknowledgments

This work was supported by the Collaborative Research Centre (SFB) 807 of the German Research Foundation (DFG) and by the Alexander von Humboldt-Foundation. We further thank Vladimir Shirokov for critical discussions.

References

1. Junge F, Haberstock S, Roos C, Stefer S, Proverbio D, Dötsch V, Bernhard F (2011) Advances in cell-free protein synthesis for the functional and structural analysis of membrane proteins. *Nat Biotechnol* 28:262–271
2. Spirin AS (2004) High-throughput cell-free systems for synthesis of functionally active proteins. *Trends Biotechnol* 22:538–545
3. Schneider B, Junge F, Shirokov VA et al (2010) Membrane protein expression in cell-free systems. *Methods Mol Biol* 601:165–186
4. Ishihara G, Goto M, Saeki M et al (2005) Expression of G protein coupled receptors in a cell-free translational system using detergents and thioredoxin-fusion vectors. *Protein Expr Purif* 41:27–37
5. Yang JP, Cirico T, Katzen F et al (2011) Cell-free synthesis of a functional G protein-coupled receptor complexed with nanometer scale bilayer discs. *BMC Biotechnol* 11:57
6. Corin K, Baaske P, Ravel DB et al (2011) Designer lipid-like peptides: a class of detergents for studying functional olfactory receptors using commercial cell-free systems. *PLoS One* 6:e25067
7. Wang X, Corin K, Baaske P et al (2011) Peptide surfactants for cell-free production of functional G protein-coupled receptors. *Proc Natl Acad Sci U S A* 108:9049–9054
8. Kaiser L, Graveland-Bikker J, Steuerwald D et al (2008) Efficient cell-free production of olfactory receptors: detergent optimization, structure, and ligand binding analyses. *Proc Natl Acad Sci U S A* 105:15726–15731
9. Ritz S, Hulko M, Zerfass C et al (2013) Cell-free expression of a mammalian olfactory receptor and unidirectional insertion into small unilamellar vesicles (SUVs). *Biochimie* 95:1909–1916
10. Sansuk K, Balog CI, van der Does AM et al (2008) GPCR proteomics: mass spectrometric and functional analysis of histamine H1 receptor after baculovirus-driven and in vitro cell free expression. *J Proteome Res* 7:621–629
11. Junge F, Luh LM, Proverbio D et al (2010) Modulation of G-protein coupled receptor sample quality by modified cell-free expression protocols: a case study of the human endothelin A receptor. *J Struct Biol* 172:94–106
12. Proverbio D, Roos C, Beyermann M et al (2013) Functional properties of cell-free expressed human endothelin A and endothelin B receptors in artificial membrane environments. *Biochim Biophys Acta* 1828:2182–2192
13. Klammt C, Schwarz D, Eifler N et al (2007) Cell-free production of G protein-coupled receptors for functional and structural studies. *J Struct Biol* 158:482–493
14. Klammt C, Perrin MH, Maslennikov I et al (2011) Polymer-based cell-free expression of ligand-binding family B G-protein coupled receptors without detergents. *Protein Sci* 20:1030–1041
15. Li Y, Wang E, Wang Y (1999) A modified procedure for fast purification of T7 RNA polymerase. *Protein Expr Purif* 16:355–358
16. Denisov IG, Grinkova YV, Lazarides AA, Sligar SG (2004) Directed self-assembly of monodisperse phospholipid bilayer nanodiscs with controlled size. *J Am Chem Soc* 126:3477–3487
17. Haberstock S, Roos C, Hoevels Y et al (2012) A systematic approach to increase the efficiency of membrane protein production in cell-free expression systems. *Protein Expr Purif* 82:308–316
18. Pédelacq JD, Cabantous S, Tran T et al (2006) Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol* 24:79–88
19. Drew D, Newstead S, Sonoda Y (2008) GFP-based optimization scheme for the overexpression and purification of eukaryotic membrane proteins in *Saccharomyces cerevisiae*. *Nat Protoc* 3:784–798
20. Banères JL, Popot JL, Mouillac B (2011) New advances in production and functional folding of G-protein-coupled receptors. *Trends Biotechnol* 29:314–322

21. Breyton C, Gabel F, Abla M et al (2009) Micellar and biochemical properties of (hemi) fluorinated surfactants are controlled by the size of the polar head. *Biophys J* 97:1077–1086
22. Lyukmanova EN, Shenkarev ZO, Khabibullina NF et al (2012) Lipid-protein nanodisks for cell-free production of integral membrane proteins in a soluble and folded state: Comparison with detergent micelles, bicelles and liposomes. *Biochim Biophys Acta* 1818:349–358
23. Roos C, Zocher M, Müller D et al (2012) Characterization of co-translationally formed nanodisc complexes with small multidrug transporters, proteorhodopsin and with the *E. coli* MraY translocase. *Biochim Biophys Acta* 1818:3098–3106
24. Dahmane T, Damian M, Mary S et al (2009) Amphipol-assisted in vitro folding of G protein-coupled receptors. *Biochemistry* 48:6516–6521
25. Basu D, Castellano JM, Thomas N, Mishra RK (2013) Cell-free protein synthesis and purification of human dopamine D2 receptor long isoform. *Biotechnol Prog* 29:601–608
26. Wang X, Cui Y, Wang J (2013) Efficient expression and immunoaffinity purification of human trace amine-associated receptor 5 from *E. coli* cell-free system. *Protein Pept Lett* 20:473–480
27. Gao T, Petrlova J, He W, Huser T et al (2012) Characterization of de novo synthesized GPCRs supported in nanolipoprotein discs. *PLoS One* 7:e44911

Chapter 11

GFP-Based Expression Screening of Membrane Proteins in Insect Cells Using the Baculovirus System

Nien-Jen Hu, Heather Rada, Nahid Rahman, Joanne E. Nettleship, Louise Bird, So Iwata, David Drew, Alexander D. Cameron, and Raymond J. Owens

Abstract

A key step in the production of recombinant membrane proteins for structural studies is the optimization of protein yield and quality. One commonly used approach is to fuse the protein to green fluorescent protein (GFP), enabling expression to be tracked without the need to purify the protein. Combining fusion to green fluorescent protein with the baculovirus expression system provides a useful platform for both screening and production of eukaryotic membrane proteins. In this chapter we describe our protocol for the expression screening of membrane proteins in insect cells using fusion to GFP as a reporter. We use both SDS-PAGE with in-gel fluorescence imaging and fluorescence-detection size-exclusion chromatography (FSEC) to screen for expression.

Key words Membrane proteins, Baculovirus, Insect cells, Green fluorescent protein, Fluorescence-detection size-exclusion chromatography

1 Introduction

The baculovirus expression system is having a major impact on the structural biology of membrane proteins. In the past 4 years, it has been used to produce most of the eukaryotic membrane proteins which have been crystallized and their structures solved by X-ray crystallography, including ion channels [1–3], G-protein-coupled receptors (GPCRs) [4–11], and transporters [12]. For GPCRs in particular, insect cells appear to be the system of choice for protein production. One of the key developments that facilitated the successful crystallization of the first GPCR, the β 2-adrenergic receptor, was insertion of T4 lysozyme into the third intracellular loop to mimic the G-protein-coupled state [13, 14]. It is now clear that irrespective of the expression technology that is used, successful production of membrane proteins for structure determination

requires the evaluation of multiple versions [15]. These include amino- and carboxy-terminal deletions (e.g., [5]), multiple orthologs to exploit natural sequence variation, and different fusion proteins [16]. Therefore, a method for expression screening membrane proteins in parallel is critical. One commonly used approach is to fuse the protein to green fluorescent protein (GFP), enabling expression to be tracked without the need to purify the protein. In this way, information about expression level, stability, and behavior in different detergents can be assessed with small amounts of unpurified material. This strategy has been adopted for screening membrane protein expression in yeast [17, 18], mammalian (e.g., human embryonic kidney cells, HEK293), and insect cells [19]. In some cases expression screening has been carried out in HEK cells and then protein production transferred to insect cells [20].

The subsequent steps in scaling up and purifying membrane proteins from insect cells are very similar for different classes of membrane protein and involve a combination of immobilized metal affinity and size-exclusion chromatography. Typically, a total membrane fraction is prepared from the insects by ultracentrifugation from which the membrane protein is extracted into detergent. However, in some reports this step has been omitted and the membrane proteins directly solubilized from broken cells [21]. The key variable is the choice of detergent for extracting and formulating the protein for which *n*-dodecyl β -D-maltoside (DDM) is most commonly used. The inclusion of cholesterol and a selective ligand (agonist or antagonist) to the detergent has been found to be important for crystallizing GPCRs [13].

In this chapter we describe our protocol for the expression screening of membrane proteins in insect cells using fusion to GFP as a reporter. We use both SDS-PAGE with in-gel fluorescence imaging and fluorescence-detection size-exclusion chromatography (FSEC) to screen for expression. The effectiveness of different detergents to solubilize the protein from isolated membranes is also evaluated using FSEC.

2 Materials

2.1 Cell Culture

1. *Spodoptera frugiperda* (Sf9) cells (ATCC® Number: CRL-1711™).
2. SF900II media containing penicillin-streptomycin (1:1,000 of 10,000 U/ml penicillin, 10,000 μ g/ml streptomycin, Life Technologies).
3. 24-Well blocks with 10 ml round-bottom wells (e.g., Qiagen 19583).
4. 24-Well tissue culture plates with lids, Erlenmeyer flask (125 ml to 2 l), and baffled Fernbach flasks (3 l).

5. 10 l Cell bag bioreactors (GE Healthcare).
6. Shaker incubator with clamps that will accommodate both flasks and 24-deep-well blocks and maintain temperature at a constant 26 °C.
7. Wave Bioreactor System 20/50 (GE Healthcare).
8. Black 50-ml falcon tubes for storing working stocks of amplified viruses (AppletonWoods BK005).

2.2 Vectors

1. Baculovirus transfer vector DNA with an A_{260}/A_{280} ratio of greater than 1.8.
2. Bacmid DNA containing the *Autographa californica* multiple nucleopolyhedrosis virus (AcMNPV) genome, e.g., Bac10:KO₁₆₂₉ [22] (see **Note 1**).

2.3 Reagents and Buffers

1. Restriction enzyme buffer (NEB buffer 3) and enzyme Bsu36I (10,000 U/ml).
2. FuGeneHD (Promega) or similar reagent optimized for transfection of insect cells.
3. Resuspension buffer: 1× PBS, pH 7.5, 0.1 mg/ml DNase I from bovine pancreas (Sigma), protease inhibitor tablet (Roche) (1 tablet in 50 ml buffer), 0.5 mM TCEP (Sigma).
4. FSEC buffer: 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.03 % DDM, 0.006 % CHS.
5. Lysis buffer: 15 mM Tris pH 7.5, 5 mM MgCl₂, 0.1 mg/ml DNase I from bovine pancreas (Sigma), protease inhibitor, 0.5 mM TCEP.
6. High salt wash buffer: 15 mM Tris pH 7.5, 1 M NaCl, 0.5 mM TCEP.
7. Storage buffer: 1× PBS pH 7.5, 100 mM NaCl, 0.5 mM TCEP.

3 Methods

3.1 Preparation of Recombinant Baculovirus

Recombinant baculoviruses are constructed by homologous recombination in insect cells between the baculovirus transfer vector and a linearized bacmid containing a disabled version of the baculovirus genome [22]. We use a transfer vector based on the pTriEx 2.0 plasmid [23] which has been engineered so that the inserted gene is expressed as an in-frame fusion with a carboxy-terminal-enhanced green fluorescent protein (GFP) gene coupled to a polyhistidine tag. Two versions of the vector are shown in Fig. 1. These pTriEx-based vectors also contain promoter elements for expression in *E. coli* and mammalian cells, enabling proteins to be tested in a multi-host screen as required.

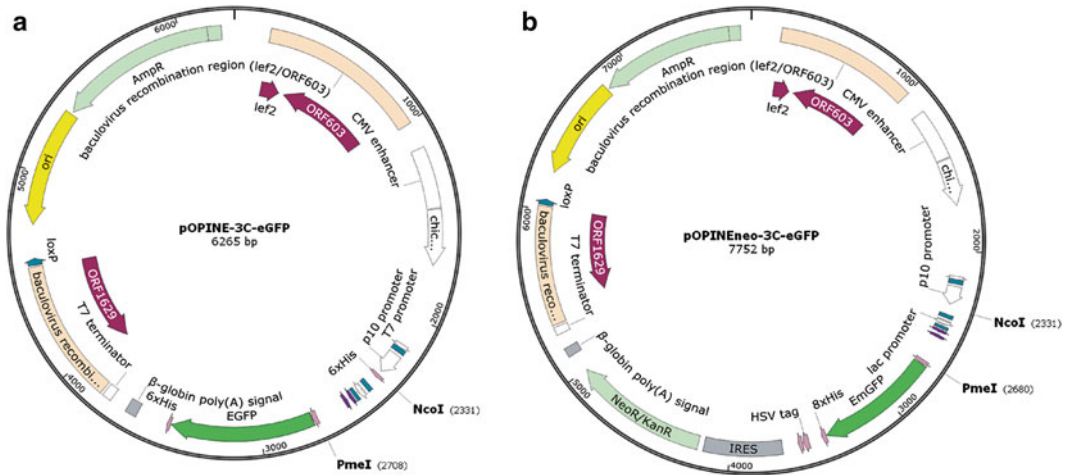


Fig. 1 Plasmid maps of baculovirus transfer vectors encoding C-terminal His-tagged GFP fusion (a) pOPIN 3C-eGFP and (b) pOPINneo 3C-eGFP. The vectors can be obtained from www.addgene.org. Plasmid maps were generated using SnapGene viewer

3.1.1 Preparation of Bacmid

1. Inoculate 200 ml of Luria Broth (LB) containing kanamycin (50 $\mu\text{g/ml}$) and chloramphenicol (30 $\mu\text{g/ml}$) with Bac10:KO₁₆₂₉ bacmid glycerol stock (*see Note 2*).
2. Grow overnight at 37 °C.
3. Purify bacmid DNA using the epicenter BacMax kit (Cambio) or equivalent.
4. Digest all of the bacmid prep in one go to maintain consistency across the prep. Dilute bacmid DNA to ~100 ng/ μl . Add NEB restriction enzyme buffer 3 and BSA to give 1 \times final concentrations.
5. Add 1 μl Bsu36I to each 6 μg bacmid and incubate for 2 h at 37 °C.
6. Add another 1 μl Bsu36I to each 6 μg bacmid and continue incubating for a further 2–3 h at 37 °C.
7. Heat at 72 °C for 20 min to stop the reaction.
8. Aliquot into usable volumes. 250 ng linearized bacmid are used per transfection so aliquots of 6 μg are useful for 24 reactions/12 μg for 48 reactions.
9. No additional purification of the bacmid is required.
10. Store the cut bacmid at –20 °C (*see Note 3*).

3.1.2 Primary Transfection to Produce P0 Virus

1. Prepare Sf9 cells in 24-well culture plates:
 - (a) Seed each well with 0.5 ml Sf900II containing Sf9 cells @ 5×10^5 cells/ml.
 - (b) Leave cells to settle for 30 min at room temperature.

2. For transfection of cells in each well, dilute 250 ng bacmid and 100–500 ng transfer vector DNA into 50 μ l Sf900II (*see Note 4*).
3. Add 1.5 μ l FuGeneHD—PIPETTE DIRECTLY INTO THE LIQUID (avoid pipetting against the plastic as this may reduce the transfection efficiency of FuGeneHD)—and mix gently. For setting up multiple reactions in parallel, make a master mix of the bacmid, FuGene, and media and add this to the aliquoted transfer vector DNA in a 96-well v-bottomed plate.
4. Incubate for 30 min at room temperature.
5. Add the transfection mix slowly (to avoid disrupting the Sf9 monolayer) to the appropriate well and gently swirl the plate to distribute the transfection mix across the well.
6. Incubate for 5–7 days at 26 °C. NB. If placing tissue culture plates in an incubator with a powerful fan, it is best to put the plates inside a plastic box along with some damp tissue to prevent the plates from drying out.
7. Harvest the supernatant from the tissue culture plates which now contains your virus. Store viral supernatant in a 96-well storage block or 1.1 ml microtubes in strips of 8. Seal with plastic lids and store at 4 °C in the dark. This is the P0 virus stock.

3.1.3 Amplification to Produce P1 and P2 Viruses

1. Prepare Sf9 cells in 24-well culture plates:
 - (a) Seed each well with 0.5 ml Sf900II containing Sf9 cells @ 1×10^6 cells/ml.
 - (b) Leave cells to settle for 30 min at room temperature.
2. Allow cells to settle for 30 min at room temperature.
3. Add 5 μ l P0 virus stock to each well (*see Note 5*).
4. Incubate for 5–7 days at 26 °C.
5. Harvest the supernatant. This is the P1 virus stock.
6. Store at 4 °C in the dark in same 96-well format as P0 virus. (If 48 or less constructs are being tested, the P0 and P1 virus stocks can be stored alongside each other in the 96-well block.)
7. For long-term storage, add serum or BSA to 10 % (v/v) final concentration and store at –80 °C.
8. Amplify viruses to produce larger working stocks by infecting 25 ml cultures of Sf9 cells (1×10^6 /ml) with 0.2 ml of P1 virus in 125 ml Erlenmeyer flask and incubating on an orbital shaker at 26 °C for 6 days.
9. Spin down cells at $1,000 \times g$ for 10 min, filter-sterilize by passing through a 0.22 μ m syringe filter, and store at 4 °C in black Falcon tubes. This is the P2 virus.

3.2 Protein Production in Baculovirus-Infected Insect Cells

Small-scale (5 ml culture volume) expression tests are carried out initially, to identify suitable constructs for further investigation. We typically screen up to 48 variants at a time which may represent different orthologs/homologs of the target membrane protein(s) and/or different carboxy-/amino-terminal truncations or in the case of GPCRs different insertion positions of, for example, T4 lysozyme. At this scale, protein expression is followed by monitoring the fluorescent signal from the GFP incorporated into the constructs (*see* Subheadings 3.4 and 3.6). For the primary screen (P1 virus), two cell/virus volume ratios are tested with cells harvested at two post-infection time points. Further optimization can be carried out in a secondary screen (P2 virus) in which viruses selected on the basis of the primary screen are titrated by expression over a longer time course, e.g., 24–96 h post-infection.

3.2.1 Small-Scale Expression Screening in 24-Deep-Well Blocks

1. Add 5 ml of Sf9 cells at 1×10^6 cells/ml to each well of a 24-deep-well block(s).
2. Add 5 or 50 μ l P1 virus to each well (*see* **Note 5**).
3. Incubate at 26 °C, with continuous shaking at 250 rpm (in an Innova42 incubator).
4. Take 1 ml samples at 48 and 72 h post-infection. Transferring samples back into a 96-well block format.
5. Harvest the cells by centrifuging the deep-well block at $6,000 \times g$ for 15 min.
6. Aspirate off the supernatant and freeze the cells at –80 °C.
7. Process the cells as described in Subheading 3.3.
8. Assess the primary expression screening results using in-gel fluorescence (*see* Subheading 3.4) and FSEC (*see* Subheading 3.5) (Fig. 2).
9. OPTIONAL: carry out a more detailed analysis of expression kinetics of selected baculoviruses which have been amplified to P2 using the protocol described above. Screen different virus/cell ratios (1:2,500 to 1:100) and infection time points (1–6 days) using cells cultured in 24-deep-well blocks. Monitor the cell viability and harvest cells from 1 ml of culture. Assess the expression parameters using in-gel fluorescence (*see* Subheading 3.4) and FSEC profiles (*see* Subheading 3.5) (Fig. 3).

3.2.2 Large-Scale Production in Shake Flasks and Cell Bag Bioreactors

Depending upon the number of infected cells that are required for downstream processing, Sf9 cells are grown in either Erlenmeyer (up to 500 ml culture volume) or Fernbach (up to 1 l culture volume) flasks or cell bag bioreactors (5 l culture volume). For membrane proteins which are expressed at relatively low levels, multi-liter culture volumes are required for purification. However, before embarking on large-scale cultures, information can be obtained at an intermediate scale (1–2 l cultures).

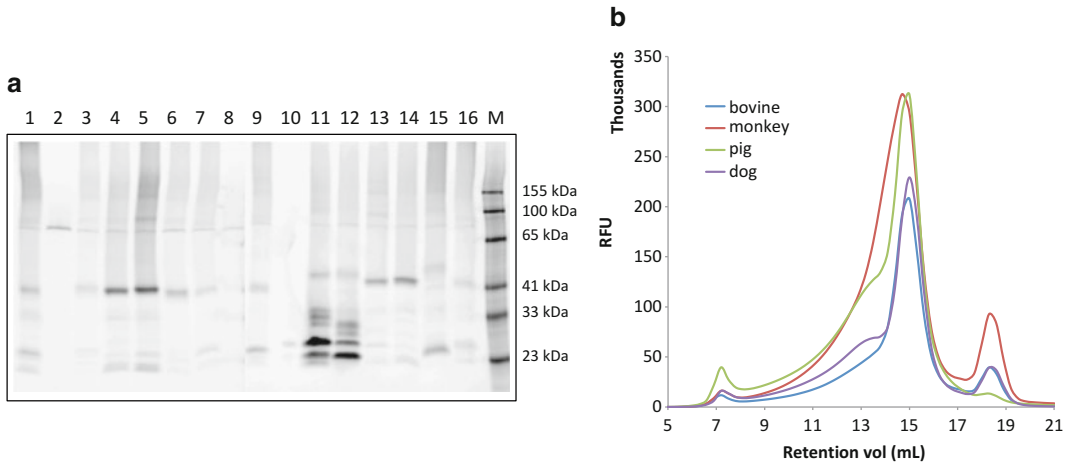


Fig. 2 Screening for expression in insect cells by in-gel fluorescence and fluorescence-detection size-exclusion chromatography (FSEC). **(a)** The secondary active transporters from 15 different vertebrate species were expressed as carboxy-terminal GFP-fusion proteins in insect cells using the baculovirus system. DDM lysates of cells were analyzed by SDS-PAGE and imaged by excitation at 488 nm and detection at 512 nm. **(b)** Samples from four species (lanes 4, 5, 13, and 14 in Fig. 2a) were analyzed further by FSEC. The fluorescence profile of each sample provides an indication of the expression level and quality of the expressed protein. In this example, all four proteins appear to be relatively monodisperse with little aggregation; only a small proportion of free GFP is present in the samples

1. Seed 2 l Erlenmeyer flasks containing 500 ml SF900II medium at 6×10^5 cells/ml (*see* **Note 6**).
2. Infect cells with ratio of P2 virus to cells (v/v) determined in the primary screen (*see* Subheading 3.2.1).
3. Follow infection of cells by assessing viability on a daily basis and harvest when cells are 75 % viable. Expression of the GFP-fusion protein in detergent-solubilized lysates of the cell samples taken each day can be followed by in-gel fluorescence and FSEC assays as described in the sections below.
4. Grow 5 l culture using the virus/cell ratio and post-infection time point determined by the intermediate-scale screening.
5. Harvest cell using centrifugation at $6,000 \times g$ for 10 min. Scrape cell pellet and transfer into a plastic bottle and store at -80°C freezer.
6. For membrane preparation from large-scale expression, refer to Subheading 3.6.

3.3 Preparation of Detergent-Solubilized Cell Lysates

1. Defrost cells, after storage at -80°C for a minimum of 20 min.
2. To a 1 ml pellet of Sf9 cells, add ice-chilled 0.2 ml resuspension buffer.
3. Shake at 900 rpm, 4°C for 10 min on a plate shaker (e.g., Vibramax 100, Heidolph).
4. Freeze samples at -80°C for 20 min.

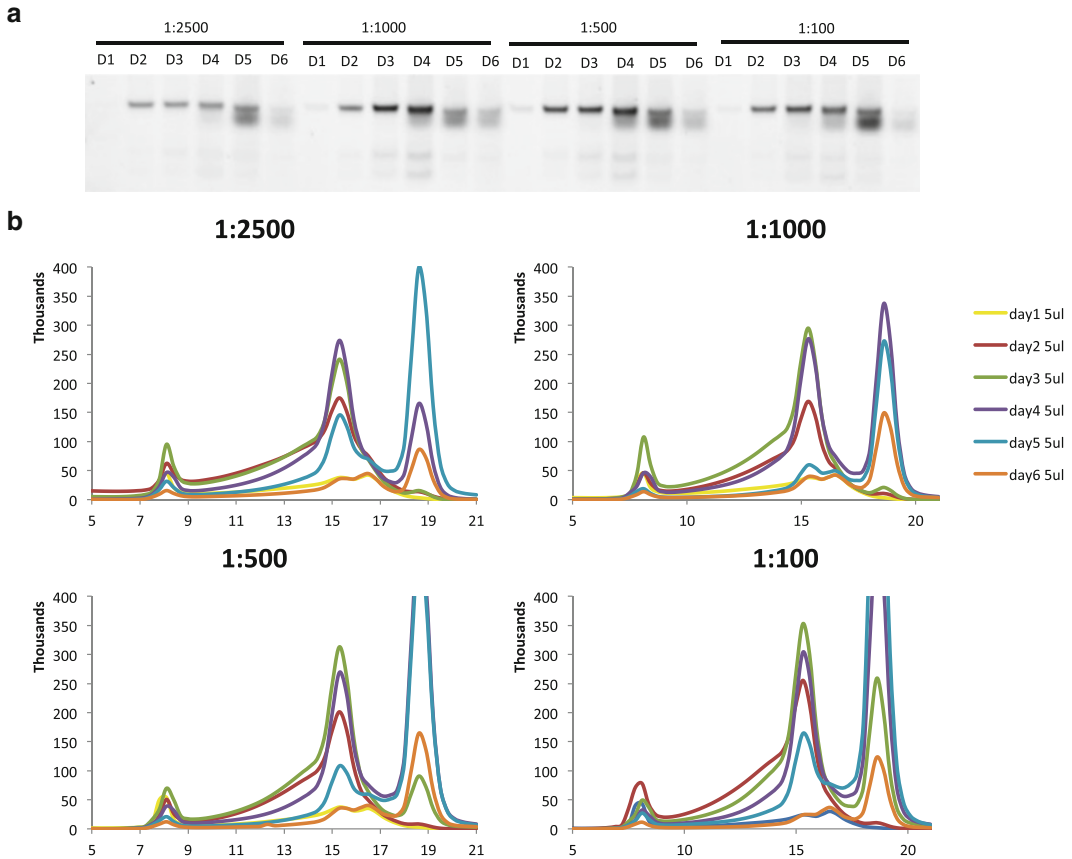


Fig. 3 Optimization of expression in a secondary screen. Expression of one of the membrane proteins from pig identified in the primary screen shown in Fig. 2 was further optimized in terms of virus/cell ratio and time of harvest post-infection. Comparing the results from the in-gel fluorescence (panel **a**) and FSEC (panel **b**) shows that infecting cells with 1:1,000 virus/cells (1×10^6 /ml) by volume and harvesting on day 3 post-infection gives the highest yield of monodisperse protein with minimal aggregate and free GFP

5. Defrost cells and transfer them to 1.5 ml Eppendorf tubes.
6. Lyse cells and extract membrane proteins by adding 10 % DDM/2 % cholesteryl hemisuccinate (CHS) mixture stock to a final concentration of 1 % DDM and 0.2 % CHS. Incubate on a rotator mixer, 4 °C for 1 h.
7. Spin sample at $150,000 \times g$ for 30 min in a bench-top ultracentrifuge to remove the detergent-insoluble fraction.
8. Collect all the supernatant with care without disturbing the pellet. The sample with an approximate volume 220 μ l is ready for the following high-throughput characterization using in-gel fluorescence (Subheading 3.4) and FSEC (Subheading 3.5).

3.4 In-Gel Fluorescence Assay

1. Mix 10 μ l of the detergent-solubilized material with 10 μ l of Novex® Tris-Glycine SDS sample buffer (2 \times). (NB. DO NOT BOIL SAMPLES as this may result in protein aggregation.)
2. Load samples onto a Novex® 12 % Tris-Glycine gel. Run the gel at 100–120 V, 4 °C, which will maintain the fused GFP tag in a folded state.
3. Use Benchmark Fluorescence marker (Invitrogen) as the protein standard (*see* **Note 7**).
4. Place the gel onto an imager with a blue light filter to detect the GFP-fusion proteins.

3.5 Fluorescence- Detection Size- Exclusion Chromatography (FSEC)

1. The high-throughput FSEC screening is performed using the Shimadzu UHPLC system equipped with an autosampler and a GFP fluorescence detector.
2. 110 μ l of solubilized samples are loaded into individual vials on the sample changer and 100 μ l of sample is injected onto a Superose 6 10/300 GL gel-filtration column pre-equilibrated with FSEC buffer.
3. The monodispersity of fluorescence profile of GFP-fusion proteins is monitored using the excitation at 488 nm and emission at 512 nm.
4. Alternatively, one can conduct FSEC using AKTA HPLC system and fractionate 0.2 ml into a black 96-well optical bottom plate in a row-by-row format. Set the measurement parameters to a 96-well fluorescence plate reader and read wells row by row. Plot the GFP fluorescence counts in each well against the fraction number [18].

3.6 Membrane Preparation from Large-Scale Expression

Assess the expression level and FSEC profiles from the primary and secondary screens (if carried out) and analyze the expression kinetics. Determine the protein target with the optimized expression conditions for the following large-scale expression. Here we present the protocol of membrane preparation from 5 l culture from cell bag bioreactor. One can adjust the buffer volume accordingly.

1. Resuspend the cell pellets harvested in Subheading 3.2.2 in 360 ml ice-chilled lysis buffer to lyse the insect cells by hypotonic “shock” (*see* **Note 8**).
2. Spin down the cells using ultracentrifugation at 150,000 g (Ti45 rotor) for 45 min.
3. Discard the supernatant and resuspend the cell pellets in 50 ml lysis buffer using Dounce homogenizer.
4. Break open the cells in the Dounce homogenizer with gentle strokes.
5. Dilute the resuspended membranes to 360 ml in low-salt lysis buffer.

6. Repeat **steps 2–5** two times.
7. Spin down the membrane fraction by ultracentrifugation at $150,000\times g$ (Ti45 rotor) for 45 min.
8. Wash the peripheral membrane proteins and ionic-bound proteins using high-salt wash buffer. Add 50 ml high-salt buffer and resuspend the membrane fractions using Dounce homogenizer with gentle strokes.
9. Dilute the resuspended membrane to 360 ml in high-salt wash buffer.
10. Spin down the membrane fractions and measure the total protein concentration of the supernatant using BCA assay.
11. Repeat **steps 8–10** several times until no obvious proteins are detectable in the supernatant.
12. Resuspend the membrane fractions in 30 ml storage buffer and measure the GFP fluorescence using a 96-well plate reader (*see Note 9*).
13. **OPTIONAL:** aliquot the 30 μ l of membranes for detergent screening. Fivefold dilute the aliquot membranes in storage buffer. To assess the monodispersity of membrane protein in different detergents, add detergent stock to final concentration of 1 % with and without 0.2 % CHS (*see Note 10*). Perform detergent solubilization and FSEC as described in Subheadings 3.3 and 3.5. Example results are shown in Fig. 4.
14. Snap freeze the membranes in liquid nitrogen and store at -80°C .
15. Perform immobilized metal affinity chromatography (IMAC) using Ni-NTA resin to purify the target protein. Use 3C protease to cleave the GFP tag and separate the cleaved GFP fusion with His tag using reverse IMAC. Use Superdex 200 column to polish the purity and monitor the monodispersity of target proteins using UV A280 detector Fig. 5b.
16. Evaluate the purity by SDS-PAGE Fig. 5a. Functional characterization should be carried out in reconstituted proteoliposome using appropriate assay.

4 Notes

1. There are a number of commercially available versions of a baculovirus bacmid which have been engineered to remove non-essential genes which may affect expression levels [24, 25].
2. If you are transforming bacmid DNA, transform into any *E. coli* cloning strain and plate out onto agar plates containing Cm only (no Kan). This is due to the transformation frequency of the large bacmid being very low.

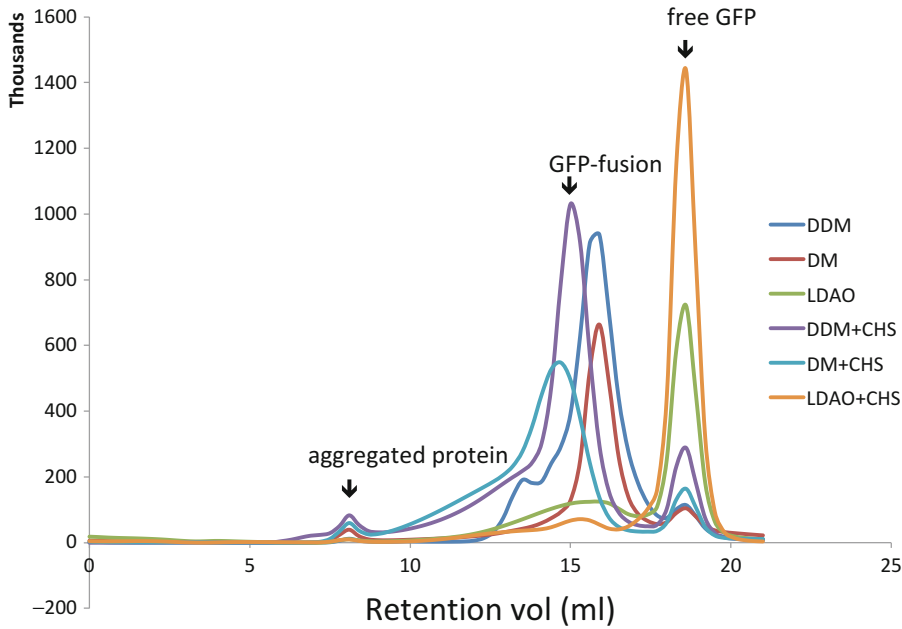


Fig. 4 Screening the effectiveness of different detergents for solubilization of the pig homolog of a secondary active transporter. Using the infection conditions identified in the secondary expression screen, transfection of Sf9 cells was scaled up as described in Subheading 3.6 and a total membrane fraction isolated from the cells. Aliquots were solubilized with different detergents (1 % v/v) and analyzed by FSEC. The results indicate that solubilization in DDM plus cholesterol (0.2 % v/v) is optimal for recovery and maintaining the stability of the protein

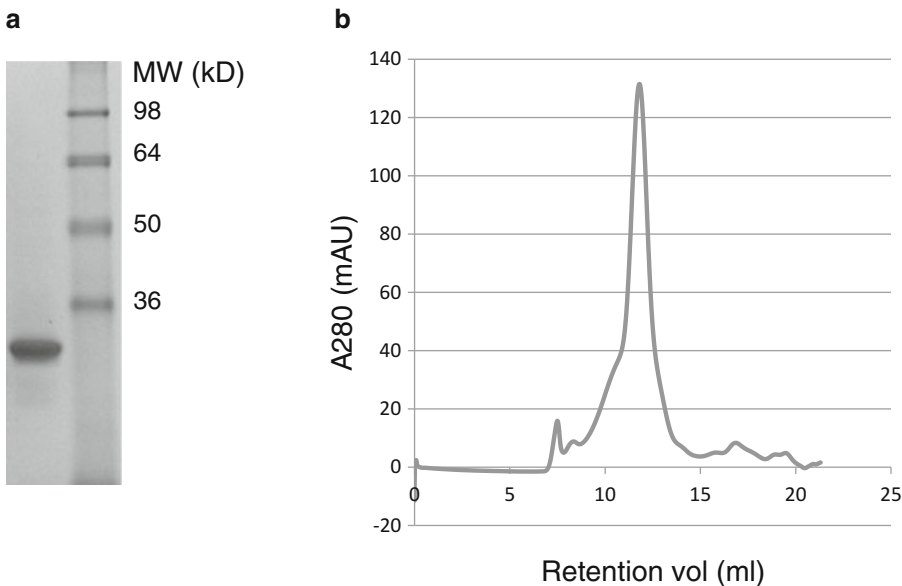


Fig. 5 Purification of the pig homolog of a secondary active transporter by a combination of Ni-chelate affinity selection and size-exclusion chromatography. (a) SDS-PAGE of purified protein stained with Coomassie blue. (b) SEC profile showing good monodispersity in the buffer containing 0.03 % DDM and 0.006 % CHS

3. The cut bacmid DNA should not be repeatedly frozen and thawed, i.e., once defrosted store in the fridge (it should be stable for 1–2 months).
4. The transfection efficiency and subsequent amplification and expression steps can be monitored by including a vector expressing GFP as a control.
5. For high-throughput addition of viruses to Sf9 cells a multi-channel, variable span pipette (like the Matrix IMPACT pipettor) can be used.
6. It is important for growing cells in shake flask cultures that the volume of media in the flask is $\leq 30\text{--}40\%$ of the flask volume to ensure adequate oxygenation of the cultures. Recently disposable 5 l flasks which can accommodate 2.5 l of culture (<http://htslabs.com>) without compromising aeration efficiency have become available, which provide a viable alternative to cell bag bioreactors.
7. As we do not completely unfold membrane proteins, they normally migrate faster on SDS-PAGE than expected, showing 70–85 % of calculated molecular weight.
8. Alternatively, insect cells can be broken using a cell disrupter (e.g., Constant Systems TS Series operated 30 psi according to the manufacturer's instructions).
9. 0.03 mg/ml of purified eGFP roughly produces 8,000 RFU counts from the 96-well microplate reader (Molecular Device M2e). Standard curve is needed for individual fluorescence spectroscopic instrument. One may have to dilute the membrane as the fluorescence signal may saturate the detection limit and thus underestimate the amount of overexpressed proteins.
10. For detergent with high critical micelle concentration (CMC) such as OG, $3\times$ CMC is higher than 1 %. Therefore, add detergent to final concentration of 2 % instead.

Acknowledgements

The OPPF-UK is funded by the Medical Research Council, UK (grant MR/K018779/1). We thank Professor Ian Jones (University of Reading) for providing the baculovirus bacmid.

References

1. Leonetti MD, Yuan P, Hsiung Y, Mackinnon R (2012) Functional and structural analysis of the human SLO3 pH- and voltage-gated K⁺ channel. *Proc Natl Acad Sci U S A* 109:19274–19279
2. Yuan P, Leonetti MD, Pico AR, Hsiung Y, MacKinnon R (2010) Structure of the human BK channel Ca²⁺-activation apparatus at 3.0 Å resolution. *Science* 329:182–186

3. Maeda S, Nakagawa S, Suga M et al (2009) Structure of the connexin 26 gap junction channel at 3.5 Å resolution. *Nature* 458:597–602
4. Granier S, Manglik A, Kruse AC et al (2012) Structure of the delta-opioid receptor bound to naltrindole. *Nature* 485:400–404
5. Hanson MA, Roth CB, Jo E et al (2012) Crystal structure of a lipid G protein-coupled receptor. *Science* 335:851–855
6. Siu FY, He M, de Graaf C, Han GW et al (2013) Structure of the human glucagon class B G-protein-coupled receptor. *Nature* 499:444–449
7. Tan Q, Zhu Y, Li J et al (2013) Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex. *Science* 341:1387–1390
8. Wacker D, Wang C, Katritch V et al (2013) Structural features for functional selectivity at serotonin receptors. *Science* 340:615–619
9. Wang C, Jiang Y, Ma J et al (2013) Structural basis for molecular recognition at serotonin receptors. *Science* 340:610–614
10. Wu B, Chien EY, Mol CD et al (2010) Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science* 330:1066–1071
11. Zhang C, Srinivasan Y, Arlow DH et al (2012) High-resolution crystal structure of human protease-activated receptor 1. *Nature* 492:387–392
12. Shintre CA, Pike AC, Li Q et al (2013) Structures of ABCB10, a human ATP-binding cassette transporter in apo- and nucleotide-bound states. *Proc Natl Acad Sci U S A* 110:9710–9715
13. Cherezov V, Rosenbaum DM, Hanson MA et al (2007) High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* 318:1258–1265
14. Rosenbaum DM, Cherezov V, Hanson MA et al (2007) GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. *Science* 318:1266–1273
15. Hanson MA, Brooun A, Baker KA et al (2007) Profiling of membrane protein variants in a baculovirus system by coupling cell-surface detection with small-scale parallel expression. *Protein Expr Purif* 56:85–92
16. Chun E, Thompson AA, Liu W et al (2012) Fusion partner toolchest for the stabilization and crystallization of G protein-coupled receptors. *Structure* 20:967–976
17. Drew D, Lerch M, Kunji E, Slotboom DJ, de Gier JW (2006) Optimization of membrane protein overexpression and purification using GFP fusions. *Nat Methods* 3:303–313
18. Newstead S, Kim H, von Heijne G, Iwata S, Drew D (2007) High-throughput fluorescent-based optimization of eukaryotic membrane protein overexpression and purification in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* 104:13936–13941
19. Kawate T, Gouaux E (2006) Fluorescence-detection size-exclusion chromatography for precrystallization screening of integral membrane proteins. *Structure* 14:673–681
20. Kawate T, Michel JC, Birdsong WT, Gouaux E (2009) Crystal structure of the ATP-gated P2X(4) ion channel in the closed state. *Nature* 460:592–598
21. Quigley A, Dong YY, Pike AC et al (2013) The structural basis of ZMPSTE24-dependent laminopathies. *Science* 339:1604–1607
22. Zhao Y, Chapman DA, Jones IM (2003) Improving baculovirus recombination. *Nucleic Acids Res* 31:E6–6
23. Bird LE, Rada H, Flanagan J et al (2014) Application of In-Fusion™ cloning for the parallel construction of *E. coli* expression vectors. *Methods Mol Biol* 1116:209–234
24. Hitchman RB, Possee RD, Siaterli E et al (2010) Improved expression of secreted and membrane-targeted proteins in insect cells. *Biotechnol Appl Biochem* 56:85–93
25. Hitchman RB, Possee RD, Crombie AT et al (2010) Genetic modification of a baculovirus vector for increased expression in insect cells. *Cell Biol Toxicol* 26:57–68

Methods for the Successful Crystallization of Membrane Proteins

Isabel Moraes and Margarida Archer

Abstract

In recent years much effort has been put towards innovative developments to overcome the numerous obstacles associated with structure determination of membrane proteins by X-ray crystallography. The advent of genomics and proteomics initiatives combined with high-throughput technologies, such as automation, miniaturization, integration, and third-generation synchrotrons, has enhanced membrane protein structure determination rate. Nevertheless, crystallization of membrane proteins still remains one of the most troublesome hurdles that every structural group must undertake. This chapter presents high-throughput methods easily available to any researcher interested in membrane protein characterization and crystallization. It is our hope this chapter can be used as a positive guide to all who are attempting crystallizing membrane proteins.

Key words Crystallization, Crystallography, Membrane proteins, Detergents, Lipidic cubic phase

1 Introduction

The study of membrane proteins is of great importance. These proteins are involved in a wide range of physiological functions. Mutations or improper folding of membrane proteins is associated with many known diseases such as heart disease, cystic fibrosis, depression, obesity, cancer, and many others [1]. The structure and function of proteins are intimately related; therefore knowledge of the three-dimensional (3D) structure of membrane proteins is key to understanding many biological processes and facilitates drug discovery programs. X-ray crystallography is still the only method capable of delivering detailed empirical information on protein structures at atomic resolution. Nevertheless, growing 3D crystals of membrane proteins which are suitable for X-ray diffraction studies is still an amazingly fine art and a major bottleneck in the field. In general, crystal formation is due to the interactions between hydrophilic regions of protein molecules. However, membrane proteins have limited hydrophilic regions,

and the presence of detergent micelles in samples of solubilized membrane proteins further reduces the number of protein–protein contacts essential for the crystal growth. Consequently, when membrane protein crystals are obtained, they are often extremely fragile with a high solvent content. Moreover, membrane proteins are very unstable and prone to aggregate which also hampers the crystallization process. In the last decade, many efforts have been made to improve the production and crystallization of membrane proteins. Recently, successful new approaches for the overexpression of membrane proteins have been developed [2–6]. Improving protein stability through mutations, deletions, monoclonal antibodies, and engineering of fusion partners have all contributed to obtaining diffraction quality crystals [7–9]. Also, new detergents and lipids have emerged as tools for the efficient solubilization and crystallization of membrane proteins [10–12]. Furthermore, crystallization trials and crystal optimization strategies have benefitted from major developments in automation and miniaturization at synchrotron beamlines [13–16].

This chapter describes three different methods with the final objective of obtaining good diffracting three-dimensional membrane protein crystals. Firstly, we focus on the importance of biophysical characterization of the protein sample prior to any crystallization attempt, as this is very important in membrane protein crystallization. Secondly, we present two different crystallization methods (vapor diffusion and lipidic cubic phase) by which membrane protein crystals can be obtained. Finally, we introduce a new method that helps researchers to improve diffraction quality of membrane protein crystals.

2 Materials

2.1 Characterization of Protein–Detergent Complexes Using a Triple Detector System Integrated with a Size-Exclusion Chromatography Column

1. Detergents: the purity of the detergents used in the protein production and crystallization should be high. In our laboratory we use detergents from Glycon and ANATRACE (ANAGRADE purity).
2. SEC column: GE Superdex 200 10/300.
3. SEC-MALLS system; Viscotek TDAmx system (Malvern) consisting of three components: (1) the GPCmax integrated solvent and sample delivery module that is equipped with an in-line de-gasser, an autosampler that can hold up to 120 vials, and an integrated pump; (2) the Triple Detector Array (TDA 305) that incorporates an RI (Refractive Index), a MALLS (Multi-Angle Laser Light Scattering), and a UV detector; all detectors are connected in series to ensure maximum sensitivity; and (3) the OmniSEC software that gives direct control of the instrument, as well as the ability to acquire and display results in real time (data acquisition rate: six channels, 5 Hz).

2.2 High-Throughput Crystallization of Membrane Proteins

2.2.1 Crystallization by Vapor Diffusion Method

1. Crystallization plates: MRC 96-well sitting drop crystallization plates (two drops) (Molecular Dimensions cat. #MD11-00-100); for the in situ experiments, hydrophobic coated CrystalQuick™ X (Greiner Bio-One) plates (Molecular Dimensions (cat. #MD11-59) or NatX-ray (cat. #609890)).
2. Crystallization commercial screens specific for membrane protein crystallization such as MemGold, MemGold2, MemStart, or MemSys from Molecular Dimensions.
3. TTP Labtech's Mosquito® Crystal protein crystallization liquid handler robot.

2.2.2 Crystallization by Lipidic Cubic Phase (LCP) Method

1. Crystallization glass sandwich plates with 200 µm spacers, the glass covers (Molecular Dimensions cat. #MD11-50-HT and cat. #MD11-52), and support frame for the robot deck (Molecular Dimensions cat. #MD11-56).
2. Crystallization commercial screens specific for membrane protein crystallization such as MemGold, MemGold2, or MemMeso from Molecular Dimensions.
3. Gastight syringes without needles (Hamilton cat. #7655-01 for the 50 µL and cat. #7656-01 for the 100 µL syringes).
4. Syringe needles (TTP Labtech cat. #4150-05902).
5. Syringes coupler purchased (TTP Labtech cat. #3072-01050).
6. Monoolein lipid (1-oleoyl-rac-glycerol) (Nu-Chek Prep).
7. TTP Labtech's Mosquito® LCP equipped with a humidifier chamber.

2.3 High-Throughput In Situ Dehydration of Membrane Protein Crystals

1. For the in situ experiments, hydrophobic coated CrystalQuick™ X (Greiner Bio-One) plates were either purchased from Molecular Dimensions (cat. #MD11-59) or NatX-ray (cat. #609890).
2. The crystallization liquid handler robot used was the TTP Labtech Mosquito® Crystal.
3. The dehydration screen was produced using a Hamilton STAR robot.

3 Methods

3.1 Characterization of Protein–Detergent Complexes Using a Triple Detector System Integrated with a Size-Exclusion Chromatography Column

Characterization of the different components present in the purified membrane protein sample is always advisable prior to setting up any crystallization trials. The information not only ensures the reproducibility of the experiments but also provides data regarding the oligomeric state and homogeneity of the solubilized membrane protein. Determination of the oligomeric state or subunit stoichiometry of membrane proteins in detergent solutions is remarkably difficult since the amount of detergent/lipid bound to the protein is unknown which in turn depends on the protein target and the

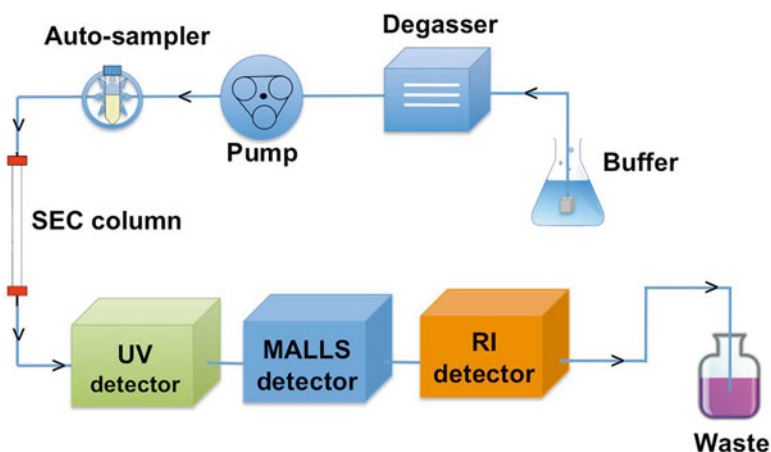


Fig. 1 Diagram of an SEC-MALLS system. The buffer is de-gassed and pumped into the system. The sample is injected into the system by an autosampler (also can be done manually) and carried through the SEC column where the size separation process takes place. When the sample elutes from the column, it passes through a series of three detectors: a UV detector, a MALLS detector, and an RI detector. The output is then analyzed by a software package (developed by the system manufacturer)

detergent used. Moreover, the quantity of free detergent in the protein sample is also unknown. Combined data from Ultraviolet (UV-280 nm), Multi-Angle Laser Light Scattering (MALLS), and Refractive Index (RI) detectors in association with a Size-Exclusion Chromatographic Column (SEC) have proven to be an accurate and reliable method to characterize membrane protein samples prior to crystallization experiments [17–19]. This technique is known as the three-detector or the SEC-MALLS method and also presents an alternative way to evaluate protein integrity and stability. The three detectors are connected in-line after a liquid chromatography system as shown in Fig. 1. Light Scattering (LS) is used to measure the absolute molecular weight of the protein sample since the overall intensity of the light scattered by the protein molecule is directly proportional to its molecular weight. The variation of the scattered light with the scattering angle is also important. Measuring the intensity of the light scattered at 90° with respect to the incident beam increases the measurement sensitivity in particular when working with protein molecules that are poor scatterers. Measuring scattering light at low-angle values is also important for accurate results when measuring large particles such as large protein complexes or aggregates. Although measurement at two different angles is sufficient, measurements at multiple angles will increase the precision and reliability of results. The UV detector measures the concentration of the protein based on absorbance at

280 nm, while the RI detector measures the protein and detergent concentration and estimates the amount of unbound detergent in the sample. The RI detector is very sensitive and is not limited to sugar-based detergents unlike other methods [18]. The value of the refractive index as a function of the concentration (dn/dc) combined with the MALLS data is used to calculate the absolute molecular weight of the membrane protein, its oligomeric state, and the size of the protein-associated detergent micelle. Please note, although dn/dc of proteins typically displays little sequence-dependent variation and therefore is “assumed” to be 0.185 for all soluble proteins, it cannot be used for Mw calculations of membrane proteins due to the presence of the detergent [19]. The dn/dc of detergents can often be obtained from the literature for most of the detergents that are used and can easily be experimentally calculated if it is unknown [20]. The use of an SEC column allows the physical separation of aggregates, protein, and free detergent micelles. It also ensures that the eluted protein sample is in the same buffer as the reference buffer (without the protein). This is mostly important for accurate determination of the baselines of the three detectors as signals are measured continuously. The SEC-MALLS method does not rely on column calibration that relates sample retention volume to molecular weight. The protocol described below provides a simple and fast mode of running an SEC-MALLS experiment.

1. Prepare the running buffer as described in **Note 1**.
2. Equilibrate the column with a minimum of 2 column volumes (CV) (*see Note 2*).
3. Calibrate the system with the protein standard (*see Note 3*).
4. Load the protein sample into the system manually or automatically (*see Note 4*).
5. Analyze results using the software provided by the manufacturer of the system (*see Note 5* and Fig. 2).

3.2 High-Throughput Crystallization of Membrane Proteins

Crystallization by vapor diffusion (in surfo) has been the most popular and successful method for crystallizing membrane proteins. This is because the method is easy to perform and identical to the one applied to soluble proteins. However, when crystallizing membrane proteins by vapor diffusion or by any other method, a large number of additional parameters need to be taken into account. Perhaps the most important variable to be considered is the presence of detergent/lipids in the protein sample. The choice of detergent is not only critical during the extraction and purification process but also important during the crystallization procedure. The shape and size of the detergent micelles plays a critical role in crystal formation. While detergents with small micelles, such as octyl- β -D-maltoside (8-carbon acyl chain) or nonyl- β -D-maltoside

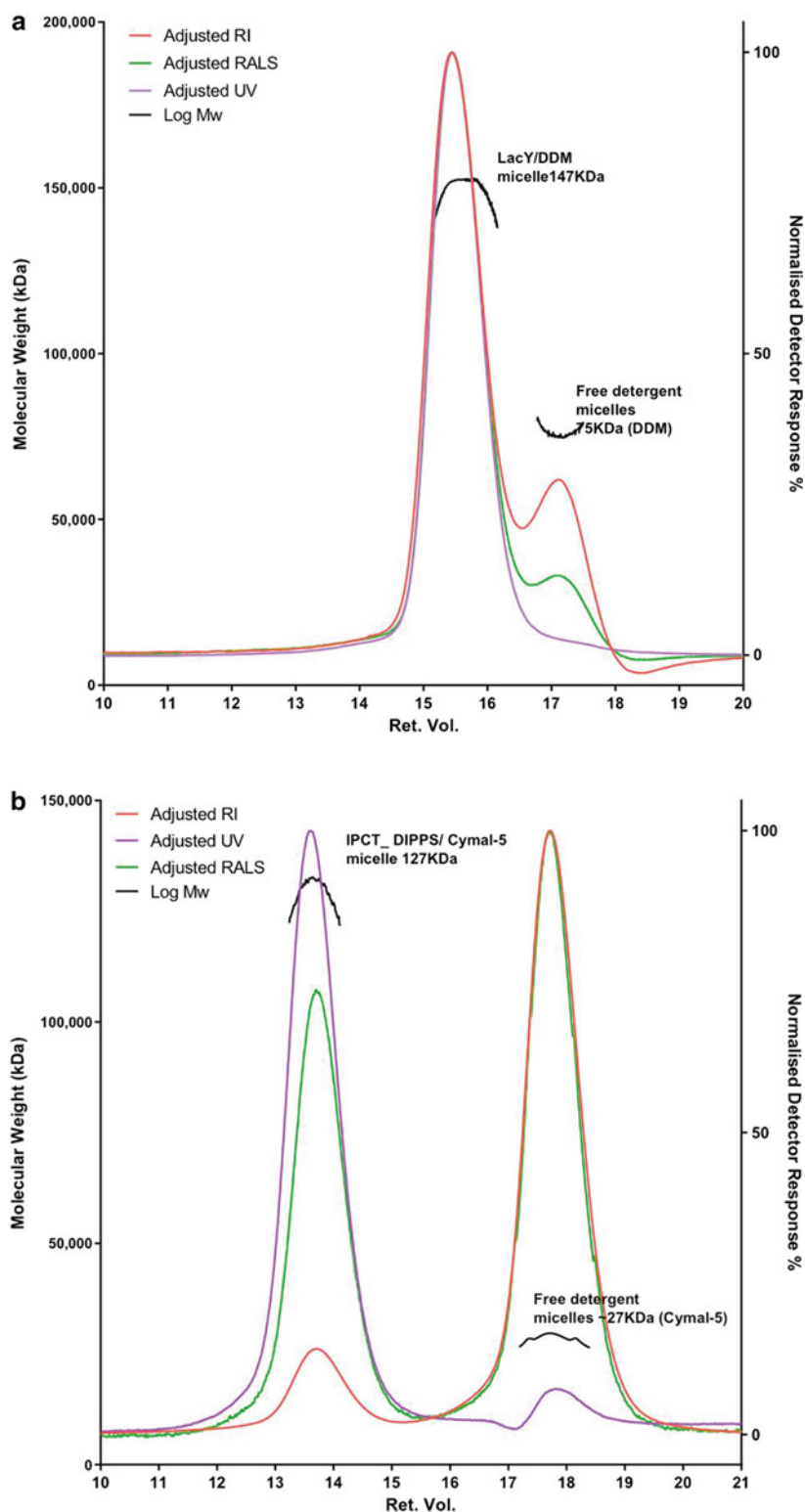


Fig. 2 SEC-MALLS analysis of two different membrane proteins (a) LacY and (b) inositol-1-phosphate cytidyltransferase (IPCT) di-myoinositol-1,3'-phosphate-1'-phosphate synthase (DIPPS). In both the cases, 160 μ L of protein sample at 1 mg/mL was injected into the system. The system was operated according to the

(9-carbon acyl chain), hardly cover the hydrophobic surface of the protein leading to protein aggregation, detergents with large micelles such as tridecyl- β -D-maltoside (13-carbon acyl chain) usually tend to engulf the entire protein [21]. Unfortunately, the choice of the best detergent for crystallization is still a trial-and-error process; hence it is recommended to start with detergents that keep the protein soluble and in its active state.

In recent years, developments in automation, miniaturization, and integration have made significant contributions to the membrane protein crystallization process. The use of liquid-handling robots has increased the number of potential crystallization conditions that can be screened while at the same time reducing the amount of protein sample required. Here, we present the two most commonly used methods for crystallizing membrane proteins using high-throughput systems.

3.2.1 Crystallization by Vapor Diffusion Method

The vapor diffusion method is based on the principle that an unsaturated concentrated protein solution is brought to supersaturation by the addition of precipitants, such as different types of polyethylene glycol (PEGs) and salts. A small droplet of protein solution is

Fig. 2 (continued) manufacturer's instructions and the analysis was performed using the OmniSEC software from Malvern. The *thick black lines* indicate the calculated molecular weight (MW) for the free detergent micelles and protein–detergent complexes. **(a)** Example of the SEC-MALLS analysis for the lactose permease (LacY) of *Escherichia coli*. The protein was purified in 0.03 % *n*-dodecyl- β -D-maltoside (DDM) by a Ni-sepharose affinity chromatography step and a final size-exclusion chromatography (SEC) step. Subsequently, the protein was concentrated using a 100-kDa centrifugal concentrator (Vivaspin). The chromatogram from the SEC-MALLS analysis shows a main peak (detected by all detectors) that corresponds to the protein complex in DDM eluting at ~15.5 mL. An additional peak at the elution volume of ~17.5 mL was observed by the RI and MALLS detectors but not by the UV detector (DDM does not absorb at 280 nm). This peak corresponds to the free detergent DDM micelles. The molecular mass of the main peak gives a value of 147 kDa corresponding to the MW of the protein in complex with the DDM detergent. The protein molecular mass was calculated to be 46 kDa, indicating that the protein was monomeric as the molecular mass of the *E. coli* LacY calculated from the protein sequence is 47 kDa. The second peak corresponding to the unbounded detergent gives a value of approximately 75 kDa, which is the molecular mass of empty DDM micelles. In conclusion, the chromatogram shows a monodisperse LacY protein sample with no significant excess free detergent. In fact the sample has produced crystals suitable for X-ray studies. **(b)** Example of the SEC-MALLS analysis for an integral membrane protein enzyme (IPCT/DIPPS) from the *Archaeoglobus fulgidus*. The protein was purified in cyclohexyl-1-pentyl- β -D-maltoside (Cymal-5) by an affinity step followed by an SEC step. The protein was concentrated using a 100-kDa centrifugal concentrator. The SEC-MALLS chromatogram shows a peak (detected by all detectors) that corresponds to the protein in complex with Cymal-5 eluting at ~14 mL. An additional peak at the elution volume of ~18 mL was observed corresponding to the free Cymal-5 micelles. The molecular mass of the main peak is 127 kDa corresponding to the MW of the protein in complex with the Cymal-5 detergent. The protein molecular mass was calculated to be 62 kDa, indicating that the protein was monomeric as the molecular mass of the IPCT/DIPPS calculated from the protein sequence is 54 kDa. The second peak corresponding to the unbound detergent is 27 kDa, which is the molecular mass of empty Cymal-5 micelles. Although the chromatogram has shown a monodisperse peak for the DIPPS protein sample, the RI peak clearly shows that the sample has excess of free Cymal-5 detergent. Despite many attempts this sample has never crystallized. In fact, the phase separation visible in most of the crystallization drops was associated with the excess of detergent in the sample

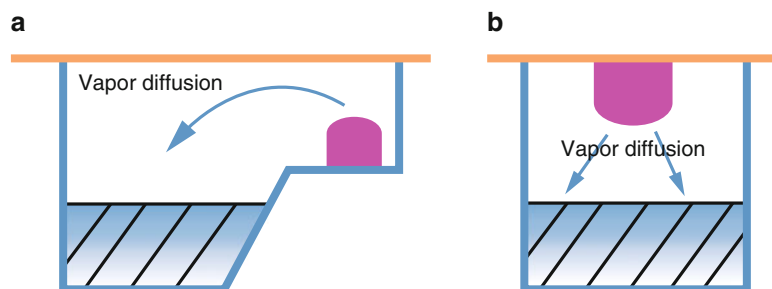


Fig. 3 Schematic representation of the sitting drop (a) and hanging drop (b) vapor diffusion method

mixed with the crystallization solution (initially at the same volume ratio) and left equilibrating over a sealed reservoir containing the same crystallization solution. Water vapor or any other volatile component evaporates from drop to the reservoir, increasing the concentration of the protein in the droplet (Fig. 3). Under the right physicochemical conditions, crystals will eventually grow. Membrane protein crystals are classified according to how crystals are formed. Type II 3D crystals are commonly observed when grown by in surfo methods. In this case, crystal packing is mostly due to interactions between hydrophilic regions of protein molecules. The presence of detergent micelles covers the hydrophobic regions of the protein which reduces the number of protein–protein contacts essential for crystal formation, resulting in extremely fragile crystals with large solvent content.

1. Concentrate the protein up to 10–12 mg/mL using the appropriate molecular weight (MW) cutoff concentrator (*see Note 6*).
2. If the protein has been frozen, centrifuge the protein sample for 15–30 min in a bench-top centrifuge (around $14,000\times g$) to remove any aggregates.
3. Pre-fill the 96-well format plates with commercially available screens specific for membrane protein crystallization [22]. The volume dispensed in the reservoir varies according to the 96-well format plate type (*see Note 7*).
4. Set up crystallization trials using a nanoliter crystallization robot. The recommended initial drop size is 100 nL (protein) + 100 nL (crystallization solution) for both sitting and hanging drop methods. Variation in the drop ratio is also advisable (*see Note 8*).
5. Store the plates at 20 °C and duplicates at 4 °C.
6. Observe the drops regularly over a period of time. The inspection of the plates can be done manually using a standard microscope or using an automated visualization system.

7. Test your crystals in a synchrotron source. DO NOT assume your crystals are protein crystals without exposing them first to the X-rays.
8. Optimize your crystals by exploring the addition of additives such as small-micelle detergents, heavy metals, salts, and organic solvents (*see* **Note 9**).

3.2.2 Crystallization by Lipidic Cubic Phase (LCP) Method

In addition to the vapor diffusion method, membrane proteins can be crystallized in lipidic cubic phase (LCP). At present, more than 50 unique membrane proteins structures have been solved using the LCP method (<http://cherezov.scripps.edu/structures.htm>). The lipidic cubic phase is spontaneously formed by gently mixing the chosen lipid with the protein–detergent complex solution at certain ratio and temperature. Lipidic cubic phases are complex three-dimensional networks of a bi-continuous lipid bilayer and two separated water channels [23]. The most common lipidic cubic phases are Im3m, Ia3d, and Pn3m [24]. For the monoolein/water system, the Pn3m phase has proven to be the most suitable for the crystallization of membrane proteins [23–28]. Crystals grown from lipidic cubic phase methods are known to be type I 3D crystals [29]. Within type I 3D crystals, proteins are organized in planar sheets through protein–detergent–lipid hydrophobic interactions stacked on top of one another by polar interactions resulting in extremely small and fragile crystals. The following procedure describes the most common way to set up the LCP crystallization using monoolein as the host lipid using a liquid-handling robot.

1. Before starting, make sure the temperature of the laboratory is set to 19–21 °C. The monoolein-based cubic phase is unstable at temperatures below 18 °C moving into lamellar phase (*see* **Note 10**).
2. Using a heating block, melt the lipid at 40 °C as the melting temperature of the monoolein is around 37 °C (*see* **Note 11**).
3. It is recommended to have the protein sample in a higher concentration than used for vapor-diffusing methods. The main reason for this is because the protein will be diluted in the lipid. It is also recommended to centrifuge the protein (around 14,000×*g*) for at least 20 min at 4 °C to remove aggregates prior to setting up the LCP experiment.
4. Using an analytical scale, tare an empty 100-μL gastight syringe (Fig. 4b).
5. Load a small amount of molten lipid (~25 mg) onto the Teflon tip of the syringe using a pipette (Fig. 4c). While dispensing the lipid with the pipette, slowly pull back the syringe plunger. If air bubbles are introduced, move the piston back and forth several times until they disappear (*see* **Note 12**).

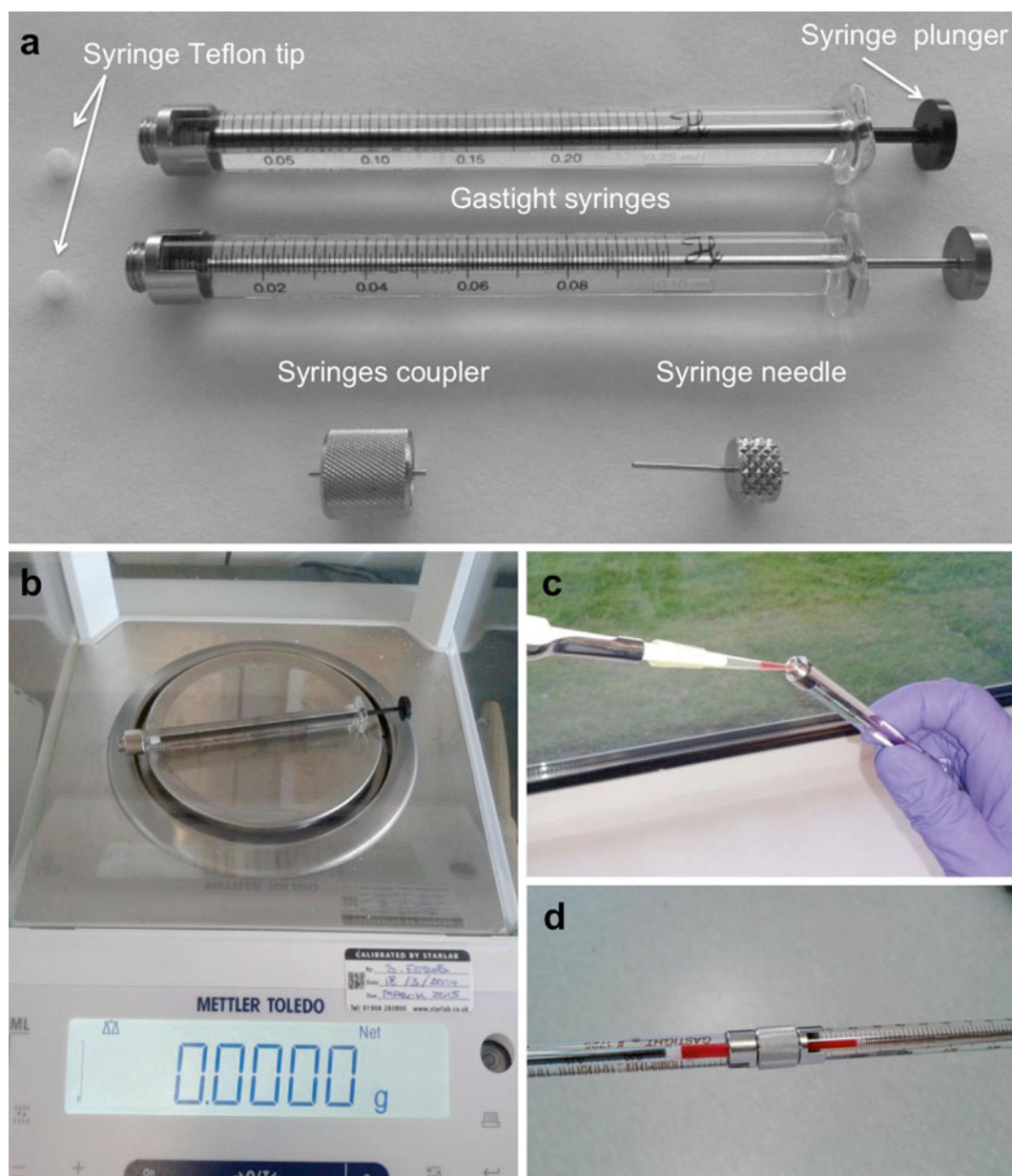


Fig. 4 Steps in the cubic mesophase preparation. **(a)** Hamilton gastight syringes and accessories for the protein and lipid mixing. **(b)** Tare the syringe that will contain the lipid. Then, the syringe will be weighted once the lipid has been loaded to determine the exact total mass of lipid in the syringe. **(c)** The lipid and the protein should be loaded onto the respective syringe through the Teflon tips with a help of a pipette. **(d)** Both syringes are coupled and ready for mixing. The protein and lipid are *red* colored for demonstration proposes

6. Weigh the syringe to determine the amount (mg) of lipid.
7. Calculate the amount of protein required using the following equation based on the monoolein/water system:

$$\text{Protein amount (mg)} = \text{monoolein (mg)} \times 2/3$$

Given that the protein solution has a similar density to the water (1.0 mg/mL), the volume of the protein solution can be directly determined from the calculated protein weight, e.g., for 25 mg of lipid, you will need around 17 μ L of protein.

8. Using a pipette, load the right amount of protein onto the Teflon tip of the 50- μ L gastight syringe avoiding air bubbles (Fig. 4c).
9. Attach the connector to the protein syringe. Push the piston slightly to remove air from the connector tube and connect the lipid syringe (Fig. 4d). Screw gently by holding the metal parts of the syringes (*see* **Note 13**).
10. Start mixing by gently pushing the PROTEIN side FIRST. The cubic phase will form spontaneously after 30–50 mixes. Vigorous mixing should be avoided, as this can cause the protein sample to heat up. Once the cubic phase is formed, it should look clear and transparent. Also, under the microscope it should be non-birefringent. Excess of protein or lipid in the mixture will result in a cloudy dispersion that will not clarify despite the number of mixes.
11. Attach the needle to the 50- μ L gastight syringe that contains the LCP phase.
12. Clamp the 50- μ L syringe to the LCP syringe dispensing pump of the Mosquito crystallization robot (Fig. 5).
13. Place the crystallization screen and the LCP glass crystallization plate on the deck of the Mosquito instrument. Remove the cover from the spacer on the LCP glass crystallization plate.
14. Operate the robot according to the manufacturer's instructions at room temperature (*see* **Note 14**).
15. In our laboratory we always use two different drop ratios (18/1 = 540 nL reservoir + 30 nL of LCP and 28/1 = 1,000 nL reservoir + 35 nL of LCP), though other ratios can also be used.
16. The process takes around 4 min to complete a single 96-well LCP plate.
17. Remove the LCP plate from the robot and seal it with the glass cover.
18. Store the plates in a temperature-controlled incubator at 20 °C and observe the drops regularly over a period of time.
19. Crystals grown in LCP are in general very small and therefore difficult to visualize. It is advisable to inspect the plates with great care using a microscope with normal light and cross-polarized light. Examples of membrane protein crystals grown in LCP are shown in Fig. 6.

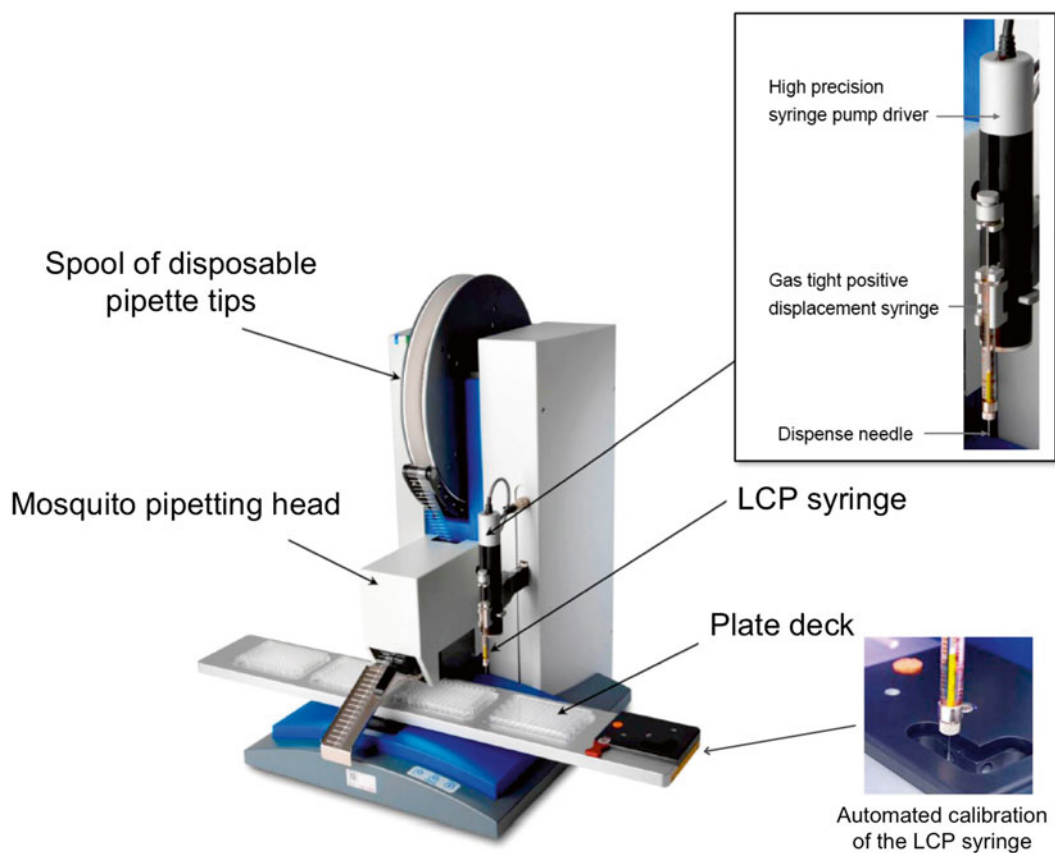


Fig. 5 The crystallization Mosquito LCP robot

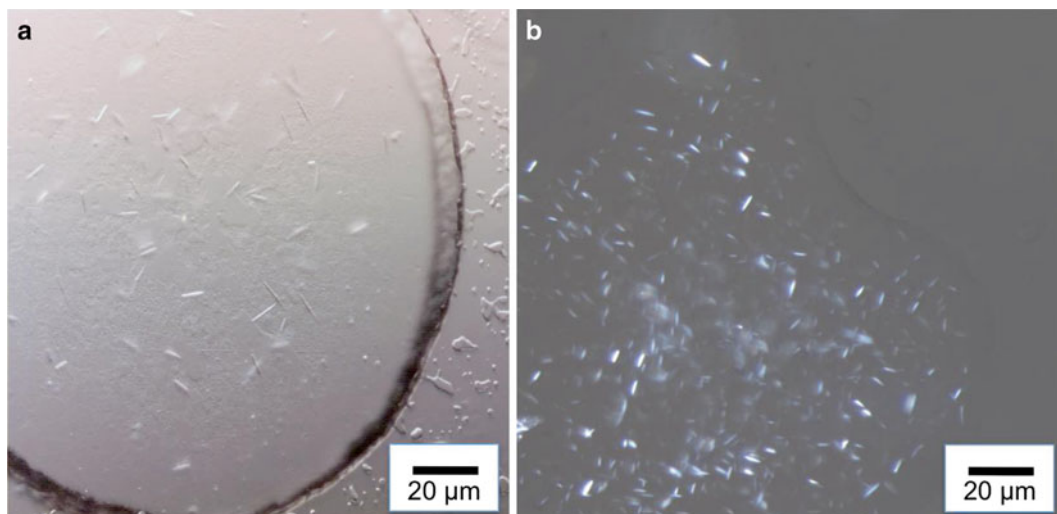


Fig. 6 Crystals of membrane proteins grown in lipidic cubic phase. (a) Crystals of a human GPCR observed with normal light. (b) Crystals of a bacterial transporter under cross-polarized light

3.3 High-Throughput In Situ Dehydration of Membrane Protein Crystals

Dehydration of protein crystals has been successfully used to improve diffraction of macromolecular crystals for many years [30–33]. Given the cost and time spent in obtaining membrane protein crystals, it is worth exploring every option available to optimize crystals; dehydration has been successful in a number of membrane proteins cases [34, 35]. As membrane protein crystals usually have a very high solvent content, successful dehydration is dependent on being able to extract only part of the available water. This may lead to new contacts between the hydrophilic areas of the proteins, potentially leading to an improved crystal lattice order. In this section we present a simple method that combines the in situ crystal dehydration with in situ plate screening to directly assess the effect of dehydration on crystal diffraction quality [14, 16, 36]. In simple terms, this method removes the critical and potentially damaging step of manually handling the crystals, providing a rapid evaluation of their quality using X-rays. The method is easily executed in the laboratory prior to be taken to an X-ray plate screening facility (in-house or at a synchrotron).

1. Using the sitting drop method (vapor diffusion), set up a 96-well crystallization plate with a nanodrop liquid handler robot (*see Note 15*). The final drop volume should be no more than 300 nL (protein sample + reservoir solution). All the 96 drops should contain the same crystallization condition, preferably the optimized condition. Allow your crystals to grow.
2. Prepare a 96 deep-well block containing increasing concentrations of different dehydrating agents (*see Note 16*). This will be your dehydration screen (Fig. 7).
3. 8–12 h before crystal diffraction screening, pierce the seal of the crystallization plate with a thin blade. Transfer the dehydration screen (using a multichannel pipette) to the wells of the crystallization plate (*see Fig. 8*).

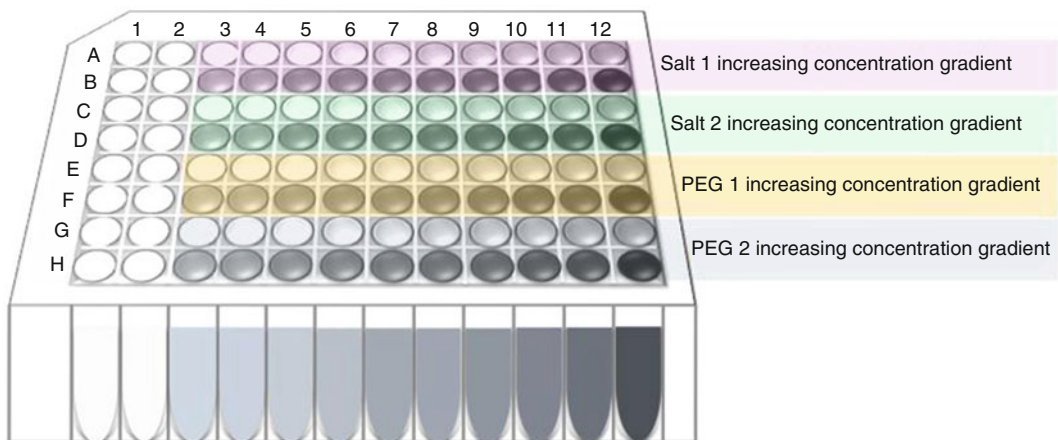


Fig. 7 The dehydration screen setup

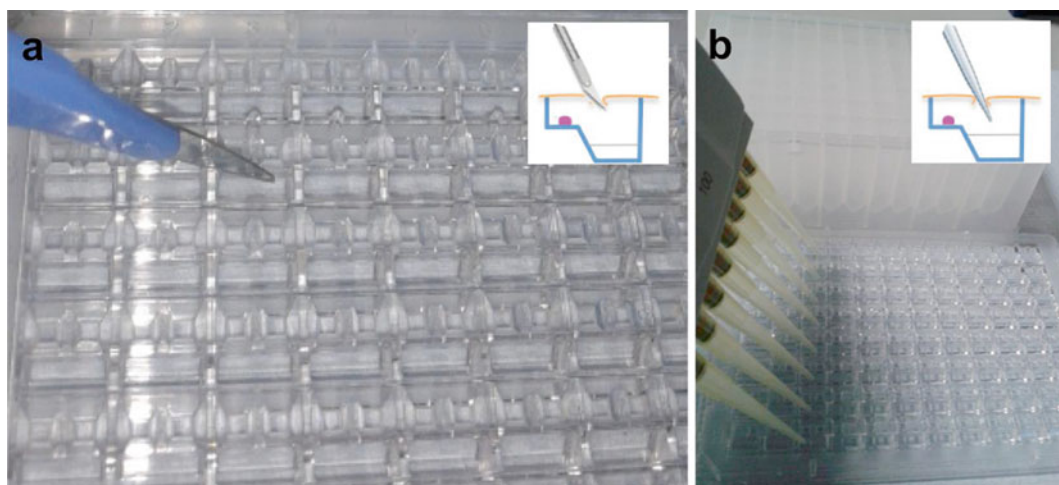


Fig. 8 Prior to data collection, (a) wells of the crystallization plate are pierced with a thin blade. (b) Dehydration solutions are then transferred from the deep-well block to the crystallization reservoir

4. In general, the volume transferred from the dehydration screen should be equal to the volume of the crystallization reservoir. Reseal the plate and leave it to equilibrate in the crystallization incubator (crystallization temperature) until beamtime.
5. The equilibrated crystallization plate is placed in the synchrotron beamline (or X-ray home source), and the crystals are tested for diffraction quality by in situ plate screening (*see Note 17 and Fig. 9*).
6. The dehydration condition that gives the best results can be reproduced at large scale (1–5 μL drops) using a 24-well plate. Dehydrated crystals from larger drops can then easily be “fished” and frozen for subsequent data collection. Please note that in general dehydrated crystals do not require addition of a cryoprotectant.

4 Notes

1. All the buffers should be prepared using HPLC-grade water or ultrapure water from a purification system. It is also important to filter the buffers through a 0.2 μm filter and de-gas them to avoid bubbles if the system in use does not have an in-line de-gasser. The presence of air bubbles in any of the detectors will greatly increase the level of baseline noise in the signal. Protein samples should also be filtered through a 0.2 μm filter or centrifuged in order to remove precipitates or any other larger insoluble particles.

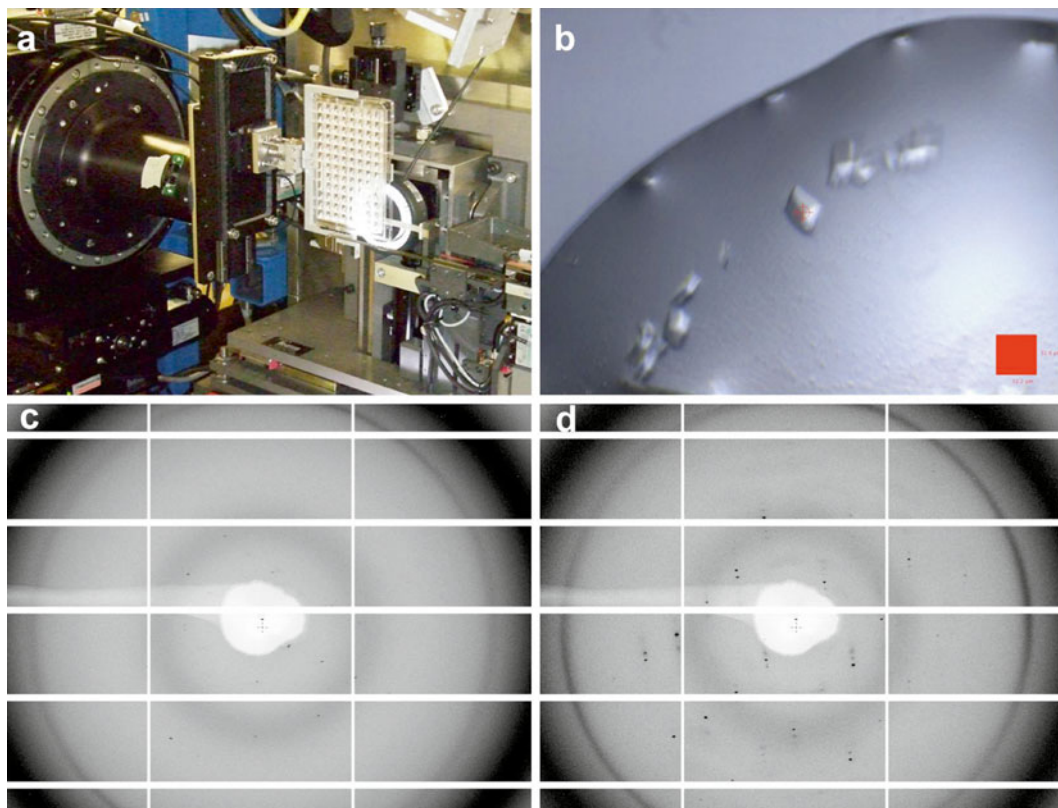


Fig. 9 In situ data collection at I24 microfocus beamline, Diamond Light Source. **(a)** In the dehydration experiment, the equilibrated crystallization plate was placed in the synchrotron beamline (I24 Diamond Light Source). **(b)** The crystals were tested for diffraction quality by in situ plate screening (room temperature). Crystals were exposed to a microfocus beam of $10 \times 10 \mu\text{m}$ and a photon flux of 2.0×10^{12} photons/s. **(c)** Diffraction pattern of the protein transporter crystals before dehydration. The best crystals have diffracted up to 9 Å. **(d)** Diffraction pattern of the crystals after dehydration. Under certain dehydration conditions, crystals have diffracted up to 6.8 Å

2. The system is extensively equilibrated with the running buffer indistinguishable from the protein buffer. Typically RI detectors are extremely sensitive, and so even small changes in running buffer versus concentration buffer can lead to errors in the results. The type of size-exclusion column integrated on the system should be appropriate for the protein sample. In general a GE Superdex 200 10/300 column (24 mL volume) is used, but for larger MW proteins or protein complexes, a GE Superose 6 10/300 is advisable. The column should be equilibrated with a minimum of 2 column volumes (CV) with a further CV through the detectors while purging the RI flow and reference cells. The column equilibration should only be complete when all detectors show stable baselines. When working with more complex buffers (e.g., additional glycerol, salts, or detergent), an increase in the equilibration time might be

needed to obtain stable baselines. It should be noted that certain types of HPLC columns are reported to “shed” (release of small particles from the stationary phase). This potentially causes interference in the detectors, increasing the signal/noise ratio. New columns should be extensively washed before used in line with the system.

3. As described in the main text, the SEC-MALLS system needs to be equilibrated prior to the loading of the protein sample. A protein standard with a well-defined dn/dc , MW, and absorbance (A_{280}) at a known concentration should be used. Bovine serum albumin (BSA) has proven to be an excellent choice for a standard. BSA is cheap and separates well under most chromatographic conditions. It is also compatible with most of the detergents, although there are suggestions that high concentrations of detergent can bind to BSA, reducing its effectiveness as a standard. The BSA with an approximate MW of 85 kDa and known dn/dc value of 0.185 is injected at 1 mg/mL (160 μ L loading volume). The resultant chromatogram is then used to calculate the detector constants for the calibration of the system for the buffer in use. The presence of detergents in membrane protein buffers has a large effect on the refractive index of the running buffer, and so this calibration helps to reduce errors by defining detector response in each set of conditions.
4. In our laboratory, 160 μ L of protein sample at 0.5–2 mg/mL is loaded onto the system by an autosampler system. Increased protein concentration leads to stronger detector signals; however the detectors can be overloaded. The experiment is performed with the column at room temperature, but because the RI is sensitive to temperature variation, all detectors are within a temperature-controlled oven that heats to 28 °C.
5. The software used in this study to acquire and analyze the data is the OmniSEC software from Viscotek (<http://www.malvern.com/en/products/product-range/viscotek-range/viscotek-systems/viscotek-tdamax/accessories/omnisec-software/>). The software imports the signals from the three different detectors and together with the calibration settings performs the data analysis. The OmniSEC software also directly controls the SEC-MALLS instrument (TDAmx system).
6. When concentrating membrane protein–detergent complexes, free detergent micelles are also concentrated if using an inappropriate molecular weight cutoff concentrator. As a result, concentrators should always have a molecular weight cutoff higher than the free detergent micelle size. Too much detergent in the protein sample can denature the protein or prevent crystal formation. It is advisable to work at detergent concentrations between two- and threefold Critical Micelle Concentration (CMC) in order to avoid excessive reduction of the protein–protein contacts essential in crystal formation.

7. The use of 96-well format plates (SBS standard) allows the screening of a large number of unique crystallization conditions using only small amounts of protein sample. These plates are compatible with all commercially available crystallization robotic systems and synchrotron beamlines. When crystallizing membrane proteins by the sitting drop method, it is recommended to use plates with round bottom wells due to the presence of detergent in the protein sample. However, if the researcher is planning to take the crystallization plates to the synchrotron for in situ crystal screening, it is advisable to use the hydrophobic plates with flat drop wells (*see* **Note 15**).
8. In our lab we set up all the crystallization trials using a crystallization robot system from TTP LabTech (Mosquito). This crystallization robot uses disposable tips to avoid cross-contamination, is easy to use, and is also fast and accurate.
9. The addition of small amphiphiles such as heptane-1,2,3-triol and benzamidine is frequently able to help as they reduce the detergent micelle size, enhancing crystal contacts [37–39].
10. According to the monoolein phase diagram, at around 20 °C the Pn3m phase is only achieved when the overall composition is about 40 % (wt/wt) protein solution. See monoolein phase diagram in Fig. 10.

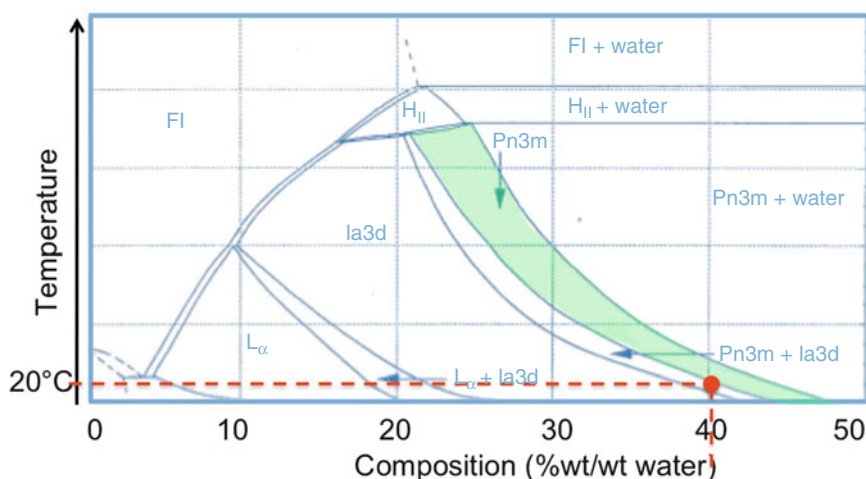


Fig. 10 Temperature-composition phase diagram of the monoolein/water system (redrawn from ref. 40). The area shaded in *green* represents the cubic phase of space group Pn3m that is associated with the membrane protein crystallization. At 20 °C and about 40 % (wt/wt) water (or protein solution), the Pn3m cubic phase forms spontaneously after mixing the monoolein with the protein solution. However, at temperatures below 17 °C, the monoolein-based cubic phase is unstable and can change into lamellar crystalline phase, damaging the protein. Keeping the temperature at 20 °C but reducing the hydration (less than 40 % wt/wt) or hydrating too much (more than 40 % wt/wt) can also lead to different phases

11. For a long-term storage, lipids should always be kept at -80°C . It is also desirable to have the lipid aliquoted into small centrifuge tubes for convenience of use. If using lipids other than monoolein, adjust the incubation temperature to a few degrees above the melting temperature of that lipid. Incubating too long may lead to degradation of the lipid.
12. The molten lipid remains liquid at room temperature for a few minutes after being taken from the heating block. This allows time to be transferred into the syringe before it solidifies. It is fine for the lipid to solidify in the syringe.
13. Do not twist excessively. This can damage the syringes' ferrules and cause leaking.
14. Although the robot operates at high speed, the use of a humidifier chamber is always advisable to avoid LCP drops drying out.
15. There are many types of SBS-format 96-well crystallization plates commercially available. However, when performing in situ plate screening, it is advisable to use crystallization plates with flat bottoms. This is because plates with round bottom wells introduce visual artifacts when mounted in the beamline. In our laboratory we always use hydrophobic-coated CrystalQuick X plates (Greiner). This plate type does not create visual artifacts and it has very low scatter background. Moreover, the hydrophobic coating of the wells improves tension surface between the protein–detergent solution and the well.
16. The dehydration screen is prepared using a liquid-handling robot. In our laboratory, we use four different types of dehydrating agents such as salts and polyethylene glycols (PEGs) at increasing concentration steps covering a large range of relative humidity. Examples of dehydrating agents include NaCl (0–5 M), LiCl (0–12 M), PEG200 (0–100 %), or ethylene glycol (0–100 %) (Fig. 7).
17. The in situ plate screening is a room temperature procedure; therefore crystals usually diffract less than when frozen. Crystal exposure time, rotation of the crystallization plate, and the number of crystals that can be tested per hour are very much dependent on the beamline specifications and the protein target. When testing crystals smaller than $20\text{ }\mu\text{m}$, it is advisable to use a microfocus X-ray beam.

Acknowledgments

The authors are grateful for the use of the Membrane Protein Laboratory funded by the Wellcome Trust (099165/Z/12/Z) at the Diamond Light Source. We also would like to thank to Mr. Mathew Jennions and Mr. James Birch for the fruitful discussions and all the MX beamline scientists at Diamond Light Source.

References

1. Sanders CR, Myers JK (2004) Disease-related misassembly of membrane proteins. *Annu Rev Biophys Biomol Struct* 33:25–51
2. Wagner S, Klepsch MM, Schlegel S et al (2008) Tuning *Escherichia coli* for membrane protein overexpression. *Proc Natl Acad Sci U S A* 105:14371–14376
3. Schlegel S, Löfblom J, Lee C et al (2012) Optimizing membrane protein overexpression in the *Escherichia coli* strain Lemo21⁺ (DE3). *J Mol Biol* 423:648–659
4. Fays FA, Zygy R-Z, Stroud RM (2010) Overexpression and purification of integral membrane proteins in yeast. *Methods Enzymol* 470:695–707
5. Drew D, Lerch M, Kunji E et al (2006) Optimization of membrane protein overexpression and purification using GFP fusions. *Nat Methods* 3:303–313
6. Tate CG (2001) Overexpression of mammalian integral membrane proteins for structural studies. *FEBS Lett* 504:94–98
7. Tate CG, Schertler GF (2009) Engineering G protein-coupled receptors to facilitate their structure determination. *Curr Opin Struct Biol* 19:386–395
8. Chun E, Thompson AA, Liu W et al (2012) Fusion partner toolchest for the stabilization and crystallization of G protein-coupled receptors. *Structure* 20:967–976
9. Rasmussen SG, Choi HJ, Fung JJ et al (2011) Structure of a nanobody-stabilized active state of the β_2 adrenoceptor. *Nature* 469:175–180
10. Privé GG (2007) Detergents for the stabilization and crystallization of membrane proteins. *Methods* 41:388–397
11. Chae PS, Rasmussen SG, Rana RR et al (2010) Maltose-neopentyl glycol (MNG) amphiphiles for solubilization, stabilization and crystallization of membrane proteins. *Nat Methods* 7:1003–1008
12. Serebryany E, Zhu GA, Yan EC (2012) Artificial membrane-like environments for in vitro studies of purified G-protein coupled receptors. *Biochim Biophys Acta* 1818:225–233
13. D'Arcy A, Villard F, Marsh M (2007) An automated microseed matrix-screening method for protein crystallization. *Acta Crystallogr D Biol Crystallogr* 63:550–554
14. Axford D, Owen RL, Foadi J et al (2012) In situ macromolecular crystallography using microbeams. *Acta Crystallogr D Biol Crystallogr* 68:592–600
15. Cherezov V, Caffrey M (2007) Miniaturization and automation for high-throughput membrane protein crystallization in lipidic mesophases. In: Chayen NE (ed) *Protein crystallization strategies for structural genomics*. International University Line, San Diego, CA
16. Moraes I, Evans G, Sanchez-Weatherby J et al (2014) Membrane protein structure determination: the next generation. *Biochim Biophys Acta* 1838:78–87
17. Wen J, Arakawa T, Philo JS (1996) Size-exclusion chromatography with on-line light-scattering, absorbance, and refractive index detectors for studying proteins and their interactions. *Anal Biochem* 240:155–166
18. Strop P, Brunger AT (2005) Refractive index-based determination of detergent concentration and its application to the study of membrane proteins. *Protein Sci* 14:2207–2221
19. Zhao H, Brown PH, Schuck P (2011) On the distribution of protein refractive index increments. *Biophys J* 100:2309–2317
20. Slotboom DJ, Duurkens RH, Olieman K, Erkens GB (2008) Static light scattering to characterize membrane proteins in detergent solution. *Methods* 46:73–82
21. Bamber L, Harding M, Monné M et al (2007) The yeast mitochondrial ADP/ATP carrier functions as a monomer in mitochondrial membranes. *Proc Natl Acad Sci U S A* 104:10830–10834
22. Newstead S, Ferrandon S, Iwata S (2008) Rationalizing α -helical membrane protein crystallization. *Protein Sci* 17:466–472
23. Landau EM, Rosenbusch JP (1996) Lipidic cubic phases: a novel concept for the crystallization of membrane proteins. *Proc Natl Acad Sci U S A* 93:14532–14535
24. Lindblom G, Rilfors L (1989) Cubic phases and isotropic structures formed by membrane lipids: possible biological relevance. *Biochim Biophys Acta* 988:221–256
25. Chiu ML, Nollert P, Loewen MEC et al (2000) Crystallization in cubo: general applicability to membrane proteins. *Acta Crystallogr D Biol Crystallogr* 56:781–784
26. Nollert P, Qiu H, Caffrey M et al (2001) Molecular mechanism for the crystallization of bacteriorhodopsin in lipidic cubic phases. *FEBS Lett* 504:179–186
27. Chung H, Caffrey M (1994) The curvature elastic-energy function of the lipid-water cubic mesophase. *Nature* 368:224–226

28. Chung H, Caffrey M (1994) The neutral area surface of the cubic mesophase: location and properties. *Biophys J* 66:377–381
29. Caffrey M, Li D, Dukkupati A (2012) Membrane protein structure determination using crystallography and lipidic mesophases: recent advances and successes. *Biochemistry* 51:6266–6288
30. Seddon JM, Templer RH, Warrender NA et al (1997) Phosphatidylcholine-fatty acid membranes: effects of headgroup hydration on the phase behaviour and structural parameters of the gel and inverse hexagonal (HII) phases. *Biochim Biophys Acta* 1327:131–147
31. Esnouf RM, Ren J, Garman EF et al (1998) Continuous and discontinuous changes in the unit cell of HIV-1 reverse transcriptase crystals on dehydration. *Acta Crystallogr D Biol Crystallogr* 54:938–953
32. Heras B, Martin JL (2005) Post-crystallization treatments for improving diffraction quality of protein crystals. *Acta Crystallogr D Biol Crystallogr* 61:1173–1180
33. Krauss IR, Sica F, Mattia CA, Merlino A (2012) Increasing the X-ray diffraction power of protein crystals by dehydration: the case of bovine serum albumin and a survey of literature data. *Int J Mol Sci* 13:3782–3800
34. Hu NJ, Iwata S, Cameron AD, Drew D (2011) Crystal structure of a bacterial homologue of the bile acid sodium symporter ASBT. *Nature* 478:408–411
35. McCusker EC, Bagn  ris C, Naylor CE et al (2012) Structure of a bacterial voltage-gated sodium channel pore reveals mechanisms of opening and closing. *Nat Commun* 3:1102
36. Douangamath A, Aller P, Lukacik P et al (2013) Using high-throughput in situ plate screening to evaluate the effect of dehydration on protein crystals. *Acta Crystallogr D Biol Crystallogr* 69:920–923
37. Timmins PA, Hauk J, Wacker T, Welte W (1991) The influence of heptane-1, 2, 3-triol on the size and shape of LDAO micelles. Implications for the crystallisation of membrane proteins. *FEBS Lett* 280:115–120
38. Schertler GF, Bartunik HD, Michel H, Oesterhelt D (1993) Orthorhombic crystal form of bacteriorhodopsin nucleated on benzamidine diffracting to 3.6   resolution. *J Mol Biol* 234:156
39. Timmins P, Pebay-Peyroula E, Welte W (1994) Detergent organisation in solutions and in crystals of membrane proteins. *Biophys Chem* 53:27–36
40. Qiu H, Caffrey M (2000) The phase diagram of the monoolein/water system: metastability and equilibrium aspects. *Biomaterials* 21: 223–234

Part IV

Structural Characterization of Proteins

Chapter 13

Application of In Situ Diffraction in High-Throughput Structure Determination Platforms

Pierre Aller, Juan Sanchez-Weatherby, James Foadi, Graeme Winter, Carina M.C. Lobley, Danny Axford, Alun W. Ashton, Domenico Bellini, Jose Brandao-Neto, Simone Culurgioni, Alice Douangamath, Ramona Duman, Gwyndaf Evans, Stuart Fisher, Ralf Flaig, David R. Hall, Petra Lukacik, Marco Mazzorana, Katherine E. McAuley, Vitaliy Mykhaylyk, Robin L. Owen, Neil G. Paterson, Pierpaolo Romano, James Sandy, Thomas Sorensen, Frank von Delft, Armin Wagner, Anna Warren, Mark Williams, David I. Stuart, and Martin A. Walsh

Abstract

Macromolecular crystallography (MX) is the most powerful technique available to structural biologists to visualize in atomic detail the macromolecular machinery of the cell. Since the emergence of structural genomics initiatives, significant advances have been made in all key steps of the structure determination process. In particular, third-generation synchrotron sources and the application of highly automated approaches to data acquisition and analysis at these facilities have been the major factors in the rate of increase of macromolecular structures determined annually. A plethora of tools are now available to users of synchrotron beamlines to enable rapid and efficient evaluation of samples, collection of the best data, and in favorable cases structure solution in near real time. Here, we provide a short overview of the emerging use of collecting X-ray diffraction data directly from the crystallization experiment. These in situ experiments are now routinely available to users at a number of synchrotron MX beamlines. A practical guide to the use of the method on the MX suite of beamlines at Diamond Light Source is given.

Key words In situ crystallography, High throughput, Automation, Ligand screening, Fragment-based drug discovery, Macromolecular crystallography, Data collection, Crystal dehydration

1 Introduction

Knowledge of the three-dimensional structures of proteins, the molecular machines of life, has transformed our understanding of all living things and has revolutionized the development of new medicines. Macromolecular crystallography (MX) provides the

most powerful tool for structural biologists to obtain high- or atomic-resolution protein structures that has made possible the elucidation of more complex and challenging targets such as macromolecular complexes, membrane proteins, and viruses [1].

Third-generation synchrotrons are now an essential part of the structure determination process and have contributed significantly to the leading role that MX occupies in structural biology. In recent years close to 90 % of all structures deposited with the Worldwide Protein Data Bank (wwPDB; wwpdb.org) have been solved from data collected at synchrotron X-ray beamlines. The rise in importance of synchrotrons for structural biology has been driven by the demands of working with small protein crystals, around 10 μm in size, which are often fragile and diffract X-rays weakly due to the high solvent content and the large unit cell dimensions of the crystals [2]. In addition, the advent of structural genomics in response to the rapid rise in the number of sequenced genomes has led to the development of high-throughput techniques for the large-scale overproduction of proteins in sufficient volumes and purity for crystallization [3]. In parallel, rapid development of hardware and software at synchrotrons to provide high-throughput data collection pipelines has been pursued to deal with the increase in demand [4]. The elimination of manual mounting of samples at synchrotron beamlines has been a fundamental part of these developments. The latter proved challenging to implement as robotics had to be developed which allowed the crystal samples to be mounted while keeping them at liquid nitrogen temperature. Two approaches in general have been developed; one involves the use of commercially available multi-axis robotic arms [5–7] and the other uses a fixed bespoke transfer mechanism [8–10]. The implementation of sample changers at beamlines quickly increased the throughput of structures determined and the rate limiting step was shifted back upstream to the laboratory and the time-consuming task of crystal harvesting from crystallization plates into standard pins for data collection [8].

An advantage of sample changers based on multi-axis robots is the ability to use different end effectors for the robotic arm. Jacquamet et al. [11] on the FIP-BM30A beamline at the ESRF exploited this to insert crystallization plates directly into the X-ray beam using a robotic arm allowing an automated scan of the 96 crystallization drops in the plate. This seminal experiment showed that diffraction could be readily recorded directly from crystallization plates and used to characterize crystals, distinguish salt crystals, and in favorable cases even collect preliminary data to check, for example, for heavy atom incorporation. The immediate advantage of such a technique was that manual handling of crystals was obviated. The disadvantage was that crystals were exposed to the bright synchrotron X-ray beams at room temperature and therefore more

susceptible to radiation damage. Consequently a large number of crystals would be required to collect a complete data set. The quality of data would also be compromised by the relatively high background generated by scatter from the plate and crystallization mother liquor. Further, the positional resolution and reproducibility of movement of the crystallization plate limited the usefulness to large crystal samples and X-ray beams of a few tens of microns to carry out the experiment. Consequently, the immediate application of the approach was seen in reducing the time and effort required to go from an initial crystallization condition to a condition producing diffraction quality crystals. So in effect, directly mounting the crystallization plate in the X-ray beam at the beamline provided a simple high-throughput and automated approach to crystal screening. Although a 96-well crystallization plate was not the ideal sample holder due to its size and large contribution to background scatter in the experiment, the method has proved extremely useful as it built on the considerable amount of automation and software development put into high-throughput crystallization.

Recently, there has been renewed interest in the use of in situ crystallization plate data collection due to the advent of fast read-out pixel array detectors [12] being coupled to intense synchrotron beamlines allowing data to be collected extremely rapidly [13, 14] making in situ data collection, in principle, tractable. Integration of a plate scanning stage with micron precision has enabled the technique to be applied to the more demanding task of characterizing smaller crystals ($<20\text{ }\mu\text{m}$) using microfocus undulator beamlines [13]. Moreover, experiments by Owen et al. [15], exploiting very high dose and frame rates, revealed a lag phase is present before radiation damage sets in. This raises the prospect of collecting complete data sets from microcrystals in the absence of cryo-cooling, re-igniting interest in data collection at room temperature. A number of de novo structures have now been published from data collected in crystallization plates in situ at synchrotron beamlines which have exemplified the merits of the technique [13, 16, 17]. Hence, as well as a fully dedicated beamline optimized to use the in situ crystallization plate approach [18], there are now a number of beamlines available for users to request the use of in situ data collection as an option (Table 1).

Alongside efforts to optimize the use of crystallization plates for diffraction data collection, ways of automating of the crystal harvesting step have been pursued. However, the challenge of integrating these methods into existing pipelines in a way that can be easily adapted for use by the large user community has proved challenging (see Deller and Rupp [19] for a detailed review).

Here, we briefly summarize practical aspects of using the in situ crystallization plate method for high-throughput screening

Table 1
In situ synchrotron beamlines worldwide

Beamline	Synchrotron	Setup	References/links
I03	DLS	Goniometer	[16]
I04-1	DLS	CATS Irelec	[21, 27]
I24	DLS	Goniometer	[13, 16, 17]
BM14	ESRF	Crystal Direct Plate screening	http://www.embl.fr/services/synchrotron_access/bm14/
FIP-BM30A	ESRF	G-rob	[11, 20]
X06DA-A	SLS	CATS Irelec	[18]
BL14.1	HZB BESSY II	CATS Irelec	[55]
Proxima 1	SOLEIL	ND	http://www.synchrotron-soleil.fr/Recherche/LignesLumiere/PROXIMA1
MX1	Australian Synchrotron	ND	http://www.synchrotron.org.au/index.php/aussyncbeamlines/macromolecular-crystallography/beamline-team
MX2	Australian Synchrotron	ND	
BL32XU	Spring8	Goniometer	[56]
W01B-MX2	LNLS	G-rob	http://lnls.cnpem.br/mx/mx2-the-beamline-2/

and evaluation of crystal samples and ligand complexes, X-ray data collection, and dehydration of crystals to improve diffraction quality using the MX beamlines at Diamond Light Source.

2 Overview of In Situ Diffraction

Eliminating the step of manual mounting has the obvious advantage that the crystals no longer need to be handled manually avoiding damage or loss of crystals. Other advantages offered by in situ diffraction [11, 13, 20] are:

1. Rapid screening of “hits” from crystallization experiments to identify false-positives such as salt or detergent crystals.
2. Rapid evaluation and characterization of crystal hits for diffraction experiments that allows ranking of samples.
3. Rapid characterization of the impact of dehydration on crystal diffraction quality.
4. Easy to use automated procedure for collection of diffraction data at room temperature.

5. High-throughput platform for ligand or fragment screening.
6. Provision of a contained sample holder that can be incorporated into workflows for working with hazard group 3 pathogens.

2.1 Plate Type

In situ plate experiments can be performed with any 96-well microtiter plate that conforms to the standards defined by the Society for Biomolecular Sciences (SBS). However, standard plates scatter X-rays significantly and usually have curved well bottoms that make centering of samples in the X-ray beam challenging and for these reasons are not recommended for screening of micron sized crystals and/or for data collection [13]. SBS plates optimized for use for X-ray diffraction have been developed. The designs of these plates have been optimized with reduced thickness of the plastic and carefully selected polymers to form the plate. A reduction in the physical profile or depth of the plate maximizes the rotation range possible during data collection. The plates currently recommended at Diamond for in situ experiments are the Greiner CrystalQuick X (Greiner Bio-One; <http://www.greinerbioone.com>) and the MiTeGen In Situ-1™ (MiTeGen; <http://mitegen.com>). Attention to the type of sealing tape used is also an important consideration and it is recommended that ThermalSeal RT for qPCR (<http://webscientific.co.uk>) be used. The scatter from these plates is compared to other frequently used plates in Fig. 1. The performance is significantly improved in the medium to low-resolution range facilitating characterization of weakly diffracting microcrystals as well as reducing the background in the higher-resolution range to extend the resolution of data that can be measured.

2.2 Transportation

Crystallization plates are surprisingly robust when it comes to transporting them between sites. As long as care is taken to maintain a suitable temperature and avoid any major physical shocks or tipping, the plates can be safely carried by ground or air transportation. Temperature stability can be recorded with USB temperature sensors which are accurate to within ± 0.5 °C and enable the effectiveness of packaging to preserve temperature during shipping to be monitored. At Diamond, crystallization plates have been delivered to the beamline by courier delivery from UK sites and plates have also been shipped successfully from the USA to the UK with no obvious detriment to the crystal's diffraction.

Some plate configurations are more suited to transportation than others. For example, microbatch experiments can be easier to transport since there is no risk of reservoir solution mixing with the crystallization drop. Low-profile crystallization plates such as the Greiner CrystalQuick X or the MiTeGen In Situ-1 plate have a dual advantage of being designed specifically for diffraction experiments and are also more compact so take up less room in the shipping container. The latter plate has micro-channels between the reservoir and protein wells, and this prevents transfer of liquid

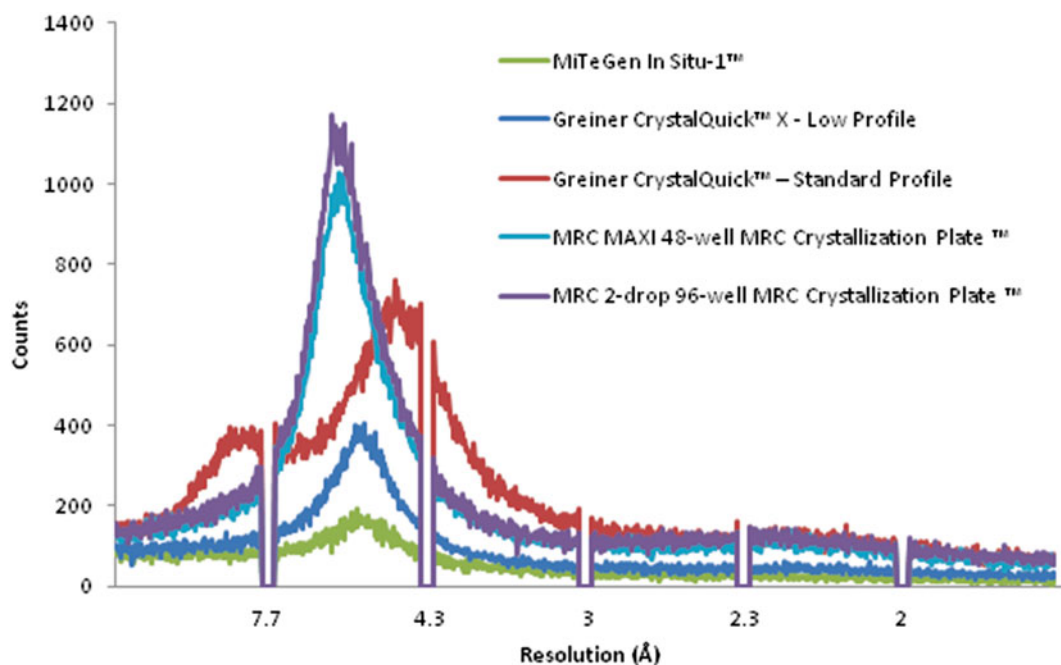


Fig. 1 Comparison of the background scatter generated by the most common 96-well microtiter plates used for crystallization and brought to Diamond for in situ diffraction experiments. The data were recorded from plates mounted on beamline I03 (342 mm plate to detector distance) exposed to a $20 \times 20 \mu\text{m}$ beam size, a wavelength of 0.9763 \AA , and a flux of 4×10^{11} photons/s. The crystallization plates specifically designed for in situ X-ray diffraction (the MiTeGen In Situ-1™ and Greiner CrystalQuick™ low profile show the lowest scatter). Gaps in graph correspond to the module gaps of the P6M Dectris PAD detector. The figure has been generated based on Bingel-Erlenmeyer et al. [18]

between the two and makes the plate more resilient to rough handling during shipping. Thus, it is highly recommended to utilize these plates for unattended transportation via courier.

Irrespective of the plate type that is used, there are general guidelines to be followed when packing the plates for shipping:

- Pack the plates into an insulated shipping container to maintain the required temperature. Include a USB temperature sensor to provide a record of temperature stability achieved during trip.
- Include gel packs that have been pre-equilibrated at the temperature of your crystallization plates.
- Add enough packing material to cushion the plates and prevent movement within the box.
- Attach labels to the outside to indicate which way is up.
- Shipments of plates containing infectious biological agents must comply with the appropriate packing requirements, e.g., biological substance, category B (UN 3373) shipping requirements. Contact your biological safety office for full details.

2.3 Beamline Setup and Access

In situ plate diffraction is currently routinely available on beamlines I03, I04-1, and I24 at Diamond Light Source with beamlines I02 and I04 soon to follow. A summary of the characteristics and capabilities of the beamlines for in situ diffraction at Diamond is summarized in Table 2. Beamline I04-1 uses the robotic arm of the CATS sample changer [6] for plate transfer and also acts as the goniometer for data collection. The impact on the experimental setup in this case is minimal and swapping between the two modes can be accomplished in less than 15 min, but requires the help of beamline staff. This allows users to work with the in situ mode and standard data collection mode as required. For example, using this mixed mode setup users can rapidly screen and rank crystal hits using an attenuated X-ray beam and harvest hits appropriately for routine cryo-cooled data collection (e.g., Fig. 2) [21]. For beamlines I03 and I24, the beamline goniometry is adapted to provide high precision plate positioning in the X-ray beams which allow microbeams to be utilized for characterization of crystals typically smaller than 20 μm [13]. These beamlines usually allocate a day a week to in situ diffraction screening experiments, and users need to request this mode of use when making a beamline application.

Figure 3 shows the setup on I03 and the main steps involved in plate transfer. The beamline has a plate hotel which can hold up to 20 plates at room temperature. The CATS or ACTOR robotic arms are used to automatically handle and transfer plates, and in the case of I04-1, the CATS robotic arm is also used to position and scan the plate in the X-ray beam.

Use of Diamond is free at the point of access for academic and peer-reviewed scientific proposals where results are to be published in peer review journals. UK and EU users are reimbursed for travel

Table 2
Diamond Light Source in situ capable beamlines summary

	I03	I04-1	I24
Setup	Goniometer	CATS Irelec Robot arm	Goniometer
Plate orientation	Portrait	Landscape	Portrait
Plate format	Greiner CrystalQuickX and MiTeGen In Situ-1™ (other SBS formats on request)	Greiner CrystalQuickX, 3 wells, low profile 96, MRC plates and MiTeGen In Situ-1™	All SBS format, including glass LCP plates
Plate hotel	5 in standard mode 20 on request	4	None
Plate mounting	Manually/robot	Robot	Manually
Oscillation range	−10° to +38°	−26° to +26°	−20° to +20°
Biocontainment	Up to level 3	Level 1	Up to level 2

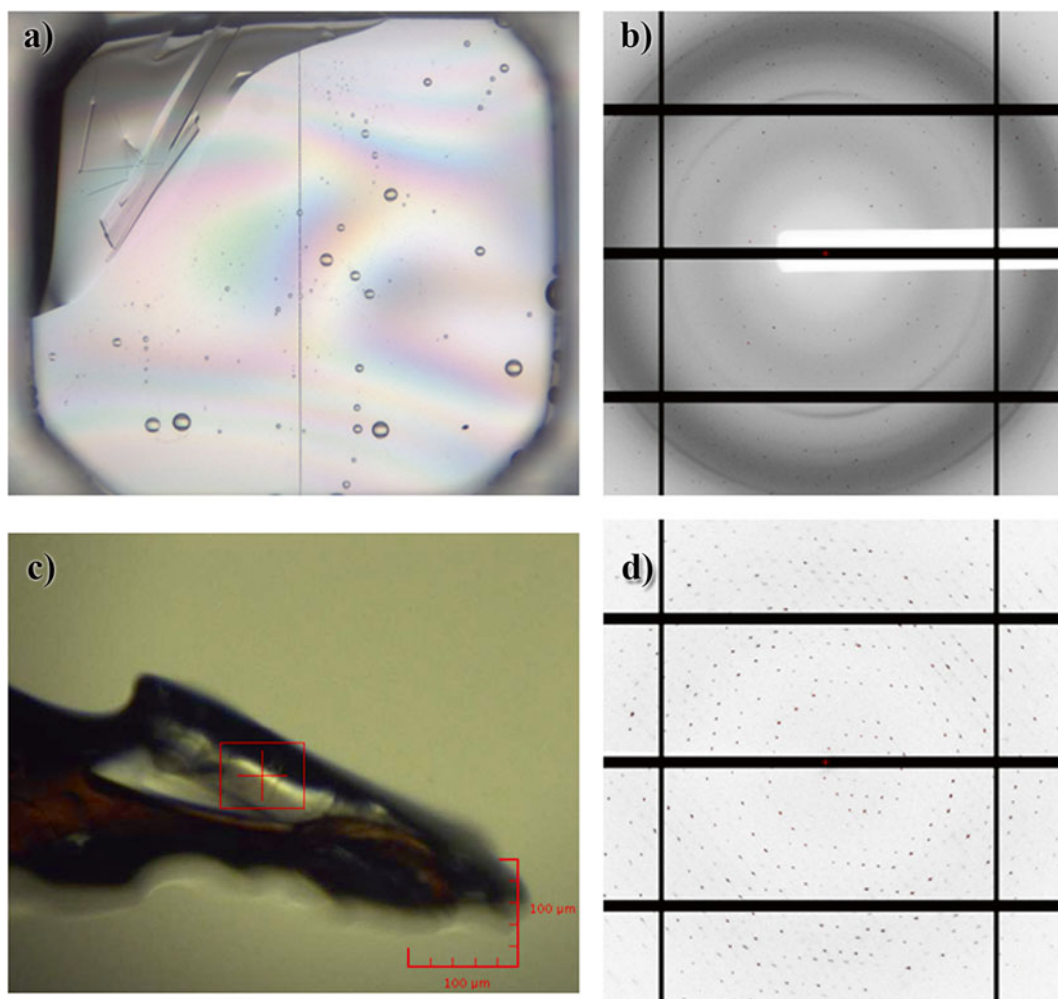


Fig. 2 Determination of the crystal structure of *Lactobacillus salivarius* UCC118 ribose 5-phosphate isomerase which was facilitated by screening of crystallization hits using in situ screening on beamline I04-1. (a) Crystal drop as viewed by the on-axis viewing system at the beamline. (b) Diffraction pattern recorded from a 4° rotation image using the in situ plate diffraction method. (c) Image of the cryo-cooled mounted crystal from the well shown in (a) mounted on the I04-1 beamline. (d) Corresponding diffraction pattern from the cryo-cooled crystal shown in (c) which diffracted 1.72 Å. Figure adapted from Lobley et al. [21]

to/from the facility and accommodation and subsistence. Users outside of the UK and Europe are reimbursed for subsistence and travel support is reviewed on a case-by-case basis.

Users can access Diamond beamlines through two main routes—the first is termed direct access and can be from either a single principal investigator which is typically for a small amount of beamtime access within a 6 month time frame or by a number of principal investigators forming a group (Block allocation Group or BAG access) which requests a significant amount of beamtime and covers access to the facility over a 2-year period. The latter model

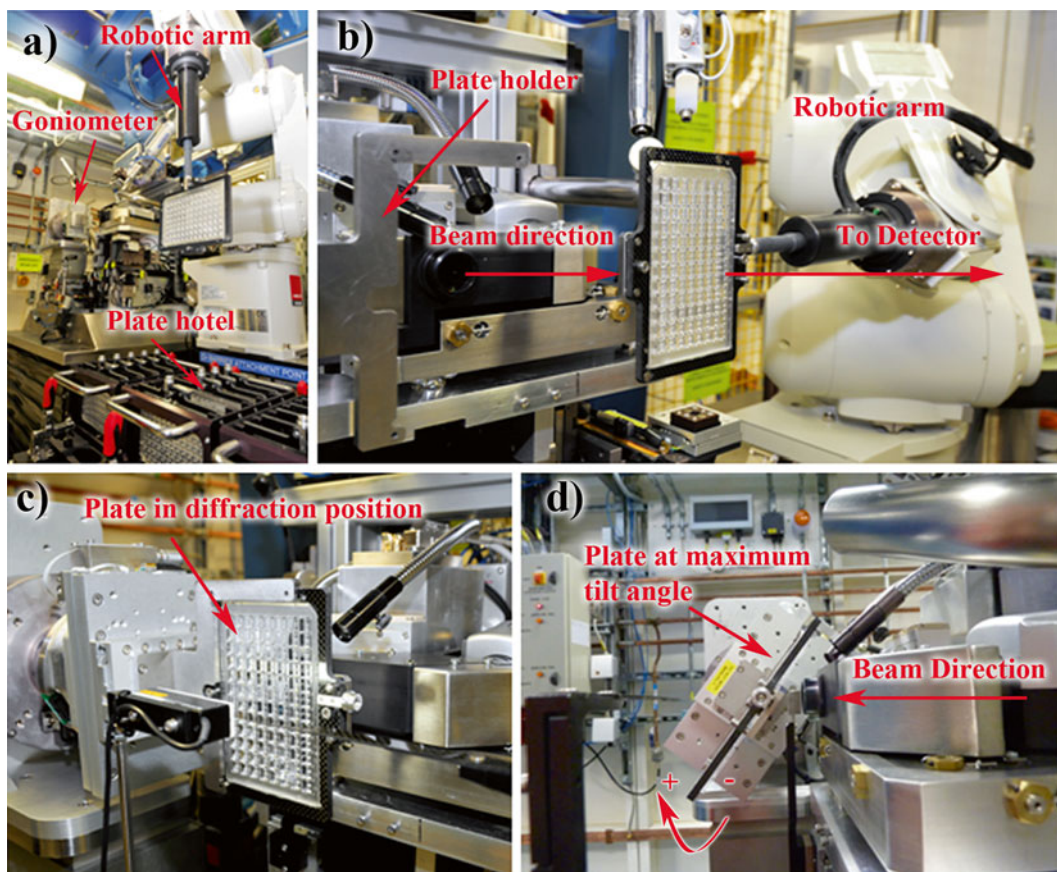


Fig. 3 In situ diffraction setup on beamline I03 at Diamond. (a) General overview of the sample environment showing the plate hotel and the Rigaku ACTOR robotic arm just after pickup of a plate from the hotel which can hold up to 20 plates. (b) The plate is inserted into a carbon frame mounted on a translation stage fixed to the beamline goniometer. (c) Detail showing the plate mounted on the goniometer, ready for data collection. (d) Close-up of the sample environment showing the plate at 38° . The total oscillation range provided for is -10° to $+38^\circ$

first implemented at the European Synchrotron Radiation Facility (ESRF) and now at a number of other European light sources provides a mechanism for frequent and regular access, and potential users are recommended to join or form such BAGS to allow efficient use of beamtime. The second route is for projects requiring access at extremely short notice and provides in principle access to beamtime within 7–10 days of requests being received (sometimes earlier depending on the request and availability of beamtime) and for obvious reasons is called “rapid access.” A complete guide and checklist of what you need to do to become a user and to obtain beamtime at Diamond can be found at the following link <http://www.diamond.ac.uk/Beamlines/Mx/Synchrotron-Access.html>.

A recipe for beamtime access is:

1. Decide on the type of experiment and requirements and choose the appropriate beamline (<http://www.diamond.ac.uk/Beamlines/Mx/>): specific issues can be discussed with the beamline scientists as required.
2. Register as a Diamond user (<http://uas.diamond.ac.uk/uas/#register>).
3. Prepare a science case to justify your request for beamtime.
4. Submit your science case and sample information for safety assessment (<http://www.diamond.ac.uk/Users/UserGuide/Proposals/Submit-Proposal.html>).
5. When you receive your beamtime allocation, provide the necessary information to gain access, validate safety, and allow accommodation and subsistence for the users attending the visit to be administered (<http://uas.diamond.ac.uk/uas/>).

2.4 Types of In Situ Plate Experiments

In situ plate diffraction experiments can be used in screening or data collection modes. In screening mode, initial crystallization hits can be rapidly assessed, as crystals do not need to be manually mounted, to distinguish protein crystals from salt or detergent crystals. This can significantly speed up optimization of crystallization conditions as manual mounting and determination of suitable cryoprotection conditions is avoided. Furthermore, as the crystals are not perturbed or modified chemically in any way, then a true assessment of the crystal diffraction quality can be ascertained. The results can then be used to prioritize and pursue optimization of conditions that gave rise to positive crystal hits, and in the case of multiple positive hits, these can be ranked and pursued appropriately based on the results of the in situ characterization. For some systems cryo-cooling of crystals disrupts, to an unacceptable level, the diffracting quality of the crystals (e.g., for whole virus crystallography), and it can frequently be difficult to find suitable cryoprotection for crystals of proteins and membrane protein crystals in particular. Therefore, facilitating the use of room temperature data collection for MX diffraction experiments is pertinent. Apart from enabling the analysis of structure at close to physiological temperatures, analyzing crystals at ambient (4–20 °C) temperatures enables the effect of modifying of the physical state of the sample in situ through dehydration to be monitored. Dehydration can improve the diffraction quality and resolution limit of a crystal. In addition the effects of chemical modification through the addition of small molecule ligands or fragments can be evaluated in, for example structure-based drug discovery.

2.4.1 In Situ Dehydration

One important parameter that affects crystal order is its hydration state. Macromolecular crystals contain between 40 % and 90 % solvent and require high relative humidity to remain stable. For example, when cryo-cooling it is important to match the osmolarity of the

crystallization condition with the cryoprotecting solution and/or limit the time spent in cryoprotectant solution. However, dehydration of macromolecular crystals can be exploited to improve the quality and diffraction limit of the crystals. Two devices have been developed that both use dew point measurements close to the sample to maintain customizable hydration states around the sample [22–24]. In the case of the HC1 device [25, 26], a large effort has been put into designing a system that is easy to use and that could be integrated into a typical beamline end station. The device which occupies a small footprint is now available for use on a number of synchrotron beamlines and its integration on Diamond beamlines allows the user to swap between using the HC1 and a standard cryostream. This has the added advantage that the user can in principle cryo-cool a hydrated crystal in situ by rapid swapping of the two devices (Fig. 4). However, dehydration experiments involve analyzing one sample at a time and require large amounts of beamtime (a typical experiment for a single sample can take up to 8 h to complete and there is no guarantee that dehydration of the crystal under study will make any improvements to the crystal diffraction quality).

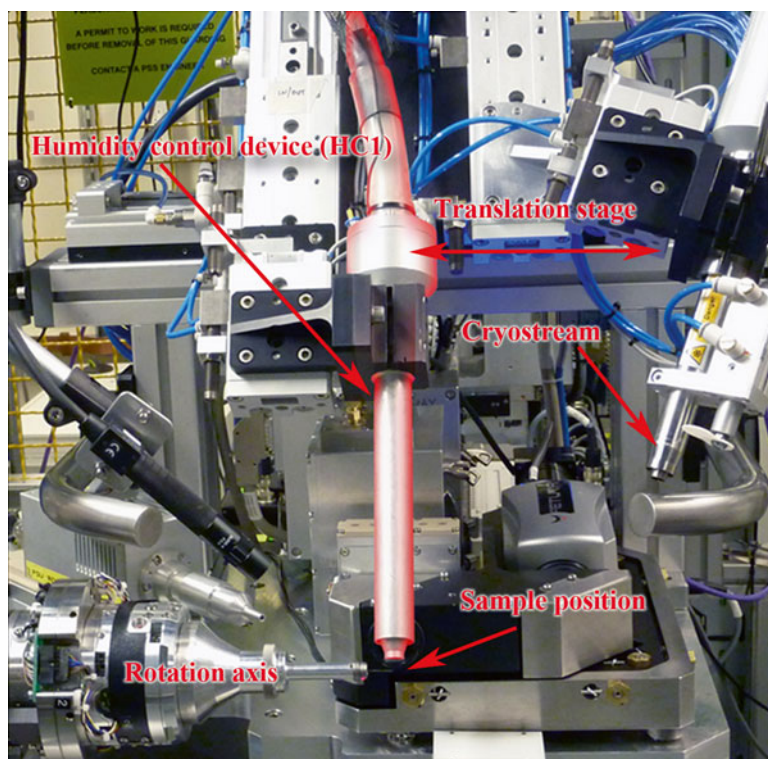


Fig. 4 Integration of the HC1 humidity controller into the end station of beamline I03 at Diamond. The HC1 and cryostream are mounted on a common translation stage allowing fast swapping between a controlled humidity environment and the standard cryo-cooled setup for data collection

Therefore, introducing a rapid way to assess the impact of dehydration is highly desirable to promote more general uptake by users. Douangamath et al. [27] have used the in situ method to analyze the effect of chemical dehydration undertaken within the crystallization plate. By simply creating a salt gradient (usually NaCl) within the plate buffer reservoirs and leaving the plate to equilibrate overnight, it is possible to dehydrate the crystals in their growth media (in situ). Using varying concentrations of salt across different wells with crystals allows screening of a range of relative humidities. For example, 0–5 M NaCl at 293 K corresponds to around 100–75 % relative humidity [28] and can be used effectively for most dehydration experiments. After 24 h of dehydration, the crystals are exposed to X-rays in situ and the dehydration effect can be directly assessed by analyzing the resulting diffraction pattern. Lattice parameters are used as indicators of the dehydration effect to identify points of interest. The main steps in the procedure are illustrated in Fig. 5. The in situ dehydration method allows rapid identification of crystal samples that can benefit from dehydration for improved diffraction. Diffraction data can be collected directly in situ or the crystal system investigated more systematically by use of the HCl device. Thus, the method provides an effective way to quickly assess the potential effects of crystal dehydration and can be easily accommodated with other high-throughput approaches, being applicable to a large number of samples at once without using an excessive amount of beamtime. The application of this protocol to membrane protein crystals is described in Chapter 13.

2.4.2 In Situ Chemical Manipulation

The protected environment of the crystallization plate can also be exploited to assess the effect of chemical additives on crystal behavior, e.g., cross-linking experiments, heavy atom soaking, varying pH, cryoprotection, and binding of inhibitors or fragments for drug design. Using essentially a similar approach to in situ dehydration, chemical additives can be easily added directly to the crystallization drop or to the well buffer depending on the type of experiment being performed. The latter has been shown to be extremely effective for well-diffracting crystals with sufficient data even being collected from a small number of crystals to allow the clear identification of bound ligand to the protein target [20].

2.5 Data Handling for In Situ Diffraction Experiments

Typically a few degrees of oscillation data can be collected from a crystal before the onset of X-ray induced radiation damage leading to partial data sets being collected from an in situ experiment at room temperature. In such cases, complete data sets for use in experimental phasing, molecular replacement, or model refinement must be assembled from many partial data sets [29]. This is not unique to in situ diffraction, as radiation damage is only partially mitigated by cryo-cooling of samples and collection of complete data sets, in particular from microcrystals, can require merging of

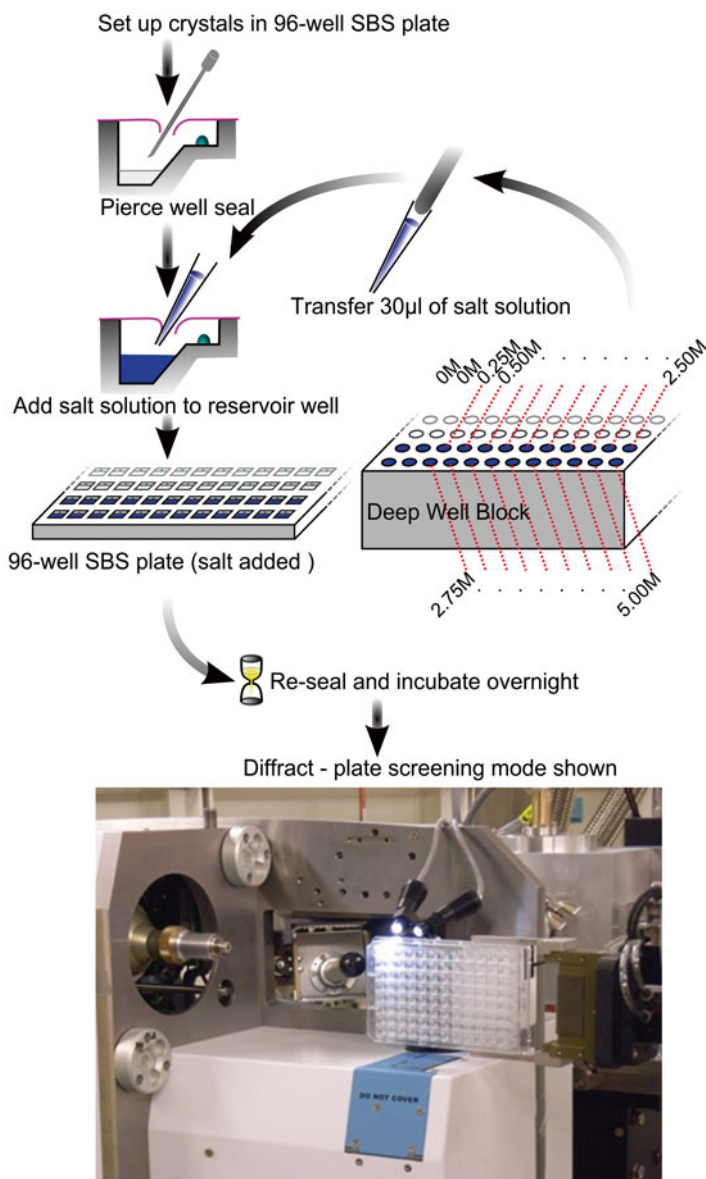


Fig. 5 Summary of the steps involved in an in situ dehydration experiment. Figure kindly reproduced with permission from Douangamath et al. [27]

multiple data sets from many samples [13]. The merging together of these partial sets can sometimes be very straightforward, but more typically it can become a tedious bookkeeping exercise that requires many scaling and merging runs in order to find the optimal combination of sufficiently isomorphous data sets. Moreover, in some special cases (such as membrane proteins), even a single crystal can exhibit “non-isomorphism.” As a result, a number of groups have developed methods to improve the management of

scaling and merging of many data sets, as well as devising new metrics to allow this to be partially or fully automated [30, 31]. At Diamond the programs xia2 [32] and *BLEND* [33] have been written to assist users in this task and provide an automated pipeline for processing of multocrystal diffraction data sets.

2.5.1 Multocrystal Data Integration with the Program xia2

xia2 is an expert system for processing primarily diffraction data from macromolecular crystals but can also be applied to data from small molecules. The program makes use of existing data reduction software including *Mosflm* [34, 35], *Labelit* [36], *AIMLESS* [37], *POINTLESS* [38], *CCP4* [39], and *XDS* [40, 43]. xia2 has been implemented at Diamond as a fully automated pipeline which requires no input from the user. The program is integrated within the beamline control graphical user interface (at Diamond the GDA [41]) which allows a fully automated data reduction pipeline to be executed. The program has been adapted to deal with multocrystal data processing which can be applied to in situ diffraction data [37]. As for standard autoprocessing with xia2, the xia2 multocrystal analysis will be automatically executed when the program detects multiple data sets recorded in a user data directory. In cases where multiple data sets are recorded into a single directory, the xia2 system at Diamond will attempt to process, scale, and merge all of the data together. The results of the analyses are presented to the user through the ISPyB database [42] and the reduced data made available in MTZ file format.

In the case of in situ diffraction data where small wedges of data are acquired, user input is often required and so it is common to run xia2 manually in this case (i.e., through the command line, though a graphical user interface is in development). For routine data sets the user needs only to guide xia2 in the pipeline to use and provide the location of the data, i.e.,

```
%xia2 (pipeline) /path/to/images
```

where (pipeline) is one of the integration options *-2d*, *-3d*, which specify integration of the data with the programs *Mosflm*/*AIMLESS* or *XDS*/*XSCALE*, respectively. A third pipeline option exists, namely, *-3dii*, which is identical to *-3d*, but uses diffraction spots from every image recorded in the data set to input for auto-indexing. This can be particularly useful for poorly diffracting or narrow sweeps of data that can prove difficult to index, which is often the case when processing in situ data, where the sweeps are typically narrow (i.e., less than 10°) and auto-indexing is challenging as one reciprocal space direction is poorly sampled. In addition, the refinement of the experimental geometry is less well constrained. Therefore, providing an accurate description of the experimental geometry and a target unit cell and symmetry can substantially improve the reliability of the processing. The command line options

```
%xia2 -cell a,b,c,al,be,ga -spacegroup SG
```

and

```
%xia2 -xparm /path/to/GXPARM.XDS
```

where GXPARM.XDS contains the refined experimental setup details that are derived from processing a relatively complete reference data set at the same distance and wavelength. The refined geometry takes into account small deviations from the ideal of the beam direction, rotation axis, and detector position. This makes the mapping of spot positions from image to reciprocal space, and hence indexing, more accurate. However, even with this additional information there may be circumstances where the data cannot be processed, e.g., poor sample alignment, or some sweeps may exhibit little or no diffraction. To increase the overall reliability when integrating a multocrystal data set, xia2 may be told to ignore processing failure of individual wedges with the failover option. Finally, the expectations encoded within xia2 for data quality are high, perhaps too high when faced with partial in situ data sets. These may be adjusted to be more lenient with the microcrystal command line option.

2.5.2 Multocrystal Scaling and Merging

The scaling and merging within xia2 for multiple data sets currently makes the assumption that the data are isomorphous and make no effort to exclude “bad” data—such a judgment is left to the user. If however the unit cell and symmetry are provided, those data successfully processed may be assumed to be approximately isomorphous. For in situ data sets, there will be cases where this assumption is false, and there is a need to select or filter the data sets to be scaled and merged together. At Diamond the program *BLEND* has been developed to achieve this task.

BLEND uses the cell dimensions of individual crystals as an indicator of isomorphism, and by determining clusters of similar crystals, it suggests several optimal merged data sets to the user. The program takes the output files from integration programs such as *Mosflm* [34, 35] and *XDS* [40, 43]. The program can be executed in three modes, *analysis*, *synthesis*, and *combination*. The first run is always in analysis mode where it examines the input data provided (i.e., the data sets collected which will consist of a set of files each containing a list of unmerged reflections from each partially recorded data set) and extracts statistical descriptors and carries out cluster analysis. The output generated is a series of ASCII files with tabulated data from all accepted sweeps, a dendrogram in both graphical and text forms, and a binary file with information needed for all runs in synthesis or combination modes. An example of this output is shown in Fig. 6. During the analysis pass the program also calculates a single parameter that provides a measure of non-isomorphism in the group of crystals investigated. This is called linear cell variation (LCV) [33] and essentially measures the largest variation of the diagonals on the three crystal cell faces, across all crystals under study. It has been observed empirically that values of

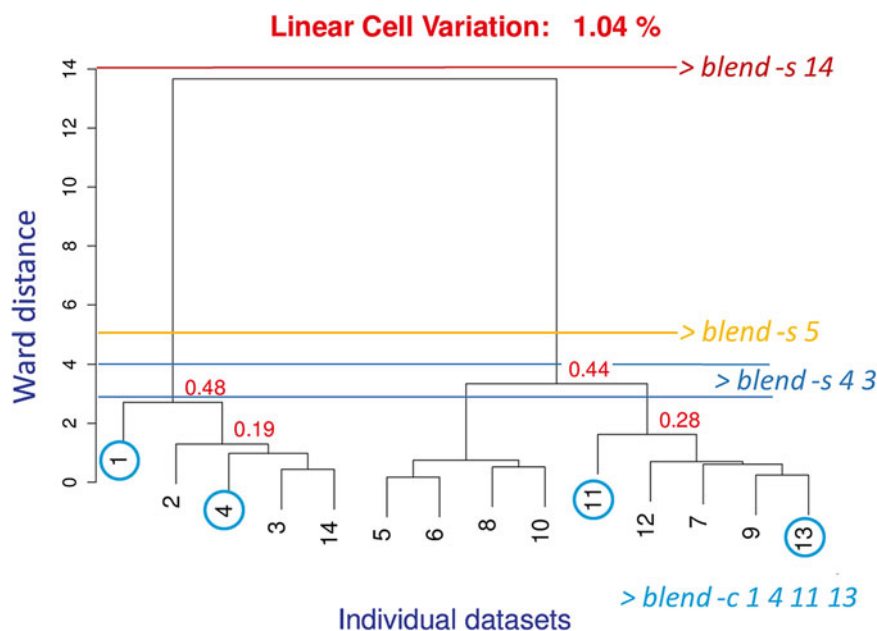


Fig. 6 An example dendrogram produced by *BLEND* following cluster analysis of 14 partial data sets. *BLEND* options for merging and combining different nodes of the dendrogram are indicated

LCV around 1.5 % or less correspond to non-noticeable structural changes.

After having executed *BLEND* in analysis mode, the user can decide whether to carry out scaling of specific clusters by providing one or two numeric values for the height in the dendrogram. This is then executed by *BLEND* in synthesis mode. More in-depth analysis of results from the synthesis run might point to specific sweeps or groups of sweeps that do not perform well but, rather, deteriorate merging statistics. In such cases it is possible to execute *BLEND* in combination mode, where sweeps to be combined do not necessarily belong to the same node of the dendrogram. Executions of the program tailored to more specific needs are provided for by simply adding or changing keywords in the input file.

Outputs from synthesis or combination runs are collected in directories “merged_files” or “combined_files,” respectively. These consist of all log files from *POINTLESS* and *AIMLESS*, mtz files for pre- and post-scaling jobs, an ASCII file with the original content of each group of sweeps, an ASCII file with overall merging statistics tabulated, and a plot of R_{meas} vs. completeness. The user can examine results from any specific group of files either by direct inspection or with *CCP4* tools like the program *LOGGRAPH* [39].

In practice the main *BLEND* steps above can be summarized in a discreet set of command line operations. To run *BLEND* analysis on a collection of data sets which are stored within a single directory,

the user only needs to input the following at the command line prompt:

```
%blend -a /where/integrated/data/are/stored
```

If the files are spread across a number of directories, then the user will have to create an ASCII file with all files (and their exact paths) listed, and this can then be read as before. For example, a file listing five data sets and called “original.dat” could contain the following data and locations:

```
/data/xtal1/x1-d01.mtz  
/data/xtal1/x1-d02.mtz  
/data/xtal3/x1-d12.mtz  
/data/xtal12/INTEGRATE.HKL  
data/xtal13/XDS_ASCII.HKL
```

The user then provides this file as the input to *BLEND* for analysis, i.e.,

```
%blend -a original.dat
```

Following analysis mode, the most relevant data output by *BLEND* is the results of the cluster analysis of all the input data sets. An example of the output of the *BLEND* analysis is shown in Fig. 6 where 14 distinct data sets were input to *BLEND* for analysis. If the number of data sets is relatively low (15–20 max), as in this case, the dendrogram can be interpreted quite easily by visual inspection and the user can decide on which data are best to merge to obtain a complete data set. This is achieved by running *BLEND* in synthesis mode. Each node in the dendrogram can give rise to a merged data set. In the case of the analysis shown in Fig. 6, to create merged files out of all nodes below height 5 in the dendrogram, the user enters

```
%blend -s 5
```

This will produce 12 new data sets: the one corresponding to node (3 + 14), the one corresponding to node (4 + 3 + 14), the one corresponding to node (2 + 4 + 3 + 14), etc. If we want to produce data sets for all nodes, we simply type

```
%blend -s 14
```

As Fig. 6 shows, cluster analysis produces a grouping of all data sets into several clusters. This does give rise to the possibility of merging various combinations of these clusters which can become challenging when many data sets are analyzed. Clustering, though, introduces limitations because the user is forced to merge data sets corresponding only to nodes in the dendrogram. In the example given in Fig. 6, it is not possible to obtain a data set out of the union of data sets 1, 4, 11 and 13 because there is no node corresponding to this union. This limitation can be overcome by running *BLEND* in combination mode as follows:

```
%blend -c 1 4 11 13
```

In summary *xia2* in concert with *BLEND* can be used in a semiautomated fashion and/or with input from the user to provide a fast and effective way of merging multiple data sets to provide the best combined data set or indeed a range of combined data sets. Developments to make the multicrystal data integration pipelines more robust and automated include:

- More reliable indexing in the case of multiple lattices on a single sweep.
- Including improved methods for resolving indexing ambiguity [44].
- Incorporation of *BLEND* cluster analyses into the *xia2* pipeline to automate merging of the best subsets of data.

3 Outlook

The use of SBS standard 96-well crystallization plates as a sample holder for diffraction data collection through the in situ method provides a seamless link between the well-developed automation systems available for crystallization and X-ray data collection. Plate designs have been optimized to reduce scatter and allow for a reasonable rotation range to be achieved using the rotation method of data collection, while retaining the SBS standard for use in high-throughput crystallization pipelines. Therefore, developing alternatives that can mimic this seamless integration is challenging. Beamline X06DA at the Swiss Light Source has a purpose-built in situ screening beamline that has an integrated crystallization facility allowing crystallization plates to be transferred in a fully automated process to screen for initial crystal hits [18]. At Diamond, a high brilliance, microfocus undulator-based beamline dedicated to in situ diffraction is currently under construction to provide a beamline that can both characterize and collect data from microcrystals. This fully automated beamline integrating crystal storage, imaging, and diffraction will come online in the autumn of 2016.

In addition to SBS standard microtiter plates for carrying out crystallization experiments, microcapillaries and chips have been developed [45–53]. A novel approach has been proposed by Cipriani et al. [54]. The method termed CrystalDirect harvests crystals from a 96-well format plate that has been modified to incorporate a thin polyimide film (Dupont™ kapton® HN general purpose film) to act as the base of the plate which can then be cut using a laser. The crystallization drops are dispensed onto this polyimide film with standard liquid handling robotics, and the drops can then be harvested in an automated fashion via a laser that the user can direct to make incisions in the polyimide foil around the identified crystal, attach it to a standard pin, and mount it on the beamline goniometer for data collection. Thus, the crystal is

harvested in a direct manner avoiding mechanical or environment stress during the mounting process. This would facilitate the harvesting of fragile difficult to mount crystals which are now frequently encountered with more complex systems under study. Although not high throughput, the method could facilitate the acceleration of these more challenging projects. In conclusion, the availability of MX beamlines offering the option of in situ crystallization plate diffraction provides users with a convenient relatively high-throughput data collection method without the need to harvest crystals. This allows collection of MX data at room temperature for a variety of experiments ranging from initial crystal hit characterization through to routine data collection as well as a platform for assessment of changes to the crystal environment such as crystal dehydration or ligand-soaking experiments.

References

- Garman EF (2014) Developments in X-ray crystallographic structure determination of biological macromolecules. *Science* 343:1102–1108
- Kantardjieff KA, Rupp B (2003) Matthews coefficient probabilities: improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Sci* 12:1865–1871
- Joachimiak A (2009) High-throughput crystallography for structural genomics. *Curr Opin Struct Biol* 19:573–584
- Beteva A, Cipriani F, Cusack S et al (2006) High-throughput sample handling and data collection at synchrotrons: embedding the ESRF into the high-throughput gene-to-structure pipeline. *Acta Crystallogr D Biol Crystallogr* 62:1162–1169
- Cohen AE, Ellis PJ, Miller MD et al (2002) An automated system to mount cryo-cooled protein crystals on a synchrotron beamline, using compact sample cassettes and a small-scale robot. *J Appl Crystallogr* 35:720–726
- Ohana J, Jacquamet L, Joly J et al (2004) CATS: a cryogenic automated transfer system installed on the beamline FIP at ESRF. *J Appl Crystallogr* 37:72–77
- <http://www.rigaku.com/products/protein/actor>
- Cipriani F, Felisaz F, Launer L et al (2006) Automation of sample mounting for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 62:1251–1259
- <http://marresearch.marxports.com/products.marxsc.html>
- Snell G, Cork C, Nordmeyer R et al (2004) Automated sample mounting and alignment system for biological crystallography at a synchrotron source. *Structure* 12:537–545
- Jacquamet L, Ohana J, Joly J et al (2004) Automated analysis of vapor diffusion crystallization drops with an X-ray beam. *Structure* 12:1219–1225
- Broennimann C, Eikenberry EF, Henrich B et al (2006) The PILATUS 1M detector. *J Synchrotron Radiat* 13:120–130
- Axford D, Owen RL, Aishima J et al (2012) In situ macromolecular crystallography using microbeams. *Acta Crystallogr D Biol Crystallogr* 68:592–600
- Owen RL, Axford D, Nettleship JE et al (2012) Outrunning free radicals in room-temperature macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 68:810–818
- Owen RL, Paterson N, Axford D et al (2014) Exploiting fast detectors to enter a new dimension in room-temperature crystallography. *Acta Crystallogr D Biol Crystallogr* 70:1248–1256
- Ren J, Wang X, Hu Z et al (2013) Picornavirus uncoating intermediate captured in atomic detail. *Nat Commun* 4:1929
- Wang X, Peng W, Ren J et al (2012) A sensor-adaptor mechanism for enterovirus uncoating from structures of EV71. *Nat Struct Mol Biol* 19:424–429
- Bingel-Erlenmeyer R, Olieric V, Grimshaw JPA et al (2011) SLS crystallization platform at beamline X06DA: a fully automated pipeline enabling in situ X-ray diffraction screening. *Cryst Growth Des* 11:916–923

19. Deller MC, Rupp B (2014) Approaches to automated protein crystal harvesting. *Acta Crystallogr F Struct Biol Commun* 70: 133–155
20. le Maire A, Gelin M, Pochet S et al (2011) In-plate protein crystallization, in situ ligand soaking and X-ray diffraction. *Acta Crystallogr D Biol Crystallogr* 67:747–755
21. Lobley CM, Aller P, Douangamath A et al (2012) Structure of ribose 5-phosphate isomerase from the probiotic bacterium *Lactobacillus salivarius* UCC118. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 68:1427–1433
22. Kiefersauer R, Stetefeld J, Gomis-Ruth FX et al (1996) Protein-crystal density by volume measurement and amino-acid analysis. *J Appl Crystallogr* 29:311–317
23. Kiefersauer R, Than ME, Dobbek H et al (2000) A novel free-mounting system for protein crystals: transformation and improvement of diffraction power by accurately controlled humidity changes. *J Appl Crystallogr* 33: 1223–1230
24. Bowler MW, Montgomery MG, Leslie AG, Walker JE (2006) Reproducible improvements in order and diffraction limit of crystals of bovine mitochondrial F(1)-ATPase by controlled dehydration. *Acta Crystallogr D Biol Crystallogr* 62:991–995
25. Russi S, Juers DH, Sanchez-Weatherby J et al (2011) Inducing phase changes in crystals of macromolecules: status and perspectives for controlled crystal dehydration. *J Struct Biol* 175:236–243
26. Sanchez-Weatherby J, Bowler MW, Huet J et al (2009) Improving diffraction by humidity control: a novel device compatible with X-ray beamlines. *Acta Crystallogr D Biol Crystallogr* 65:1237–1246
27. Douangamath A, Aller P, Lukacik P et al (2013) Using high-throughput in situ plate screening to evaluate the effect of dehydration on protein crystals. *Acta Crystallogr D Biol Crystallogr* 69:920–923
28. Winston PW, Bates DH (1960) Saturated solutions for the control of humidity in biological research. *Ecology* 41:232–237
29. Ji X, Sutton G, Evans G et al (2010) How baculovirus polyhedra fit square pegs into round holes to robustly package viruses. *EMBO J* 29:505–514
30. Giordano R, Leal RMF, Bourenkov GP et al (2012) The application of hierarchical cluster analysis to the selection of isomorphous crystals. *Acta Crystallogr D Biol Crystallogr* 68:649–658
31. Liu Q, Dahmane T, Zhang Z et al (2012) Structures from anomalous diffraction of native biological macromolecules. *Science* 336: 1033–1037
32. Winter G (2010) xia2: an expert system for macromolecular crystallography data reduction. *J Appl Crystallogr* 43:186–190
33. Foadi J, Aller P, Alguel Y et al (2013) Clustering procedures for the optimal selection of data sets from multiple crystals in macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 69:1617–1632
34. Leslie AG (2006) The integration of macromolecular diffraction data. *Acta Crystallogr D Biol Crystallogr* 62:48–57
35. Leslie AG, Powell HR (2007) Processing diffraction data with mosflm evolving. *Methods Macromol Crystallogr* 245:41–51
36. Sauter NK, Grosse-Kunstleve RW, Adams PD (2004) Robust indexing for automatic data collection. *J Appl Crystallogr* 37:399–409
37. Evans PR, Murshudov GN (2013) How good are my data and what is the resolution? *Acta Crystallogr D Biol Crystallogr* 69:1204–1214
38. Evans P (2011) An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallogr D Biol Crystallogr* 6:282–292
39. Winn MD, Ballard CC, Cowtan KD et al (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* 67:235–242
40. Kabsch W (2010) XDS. *Acta Crystallogr D Biol Crystallogr* 66:125–132
41. <http://www.opengda.org/>
42. Delageniere S, Brenchereau P, Launer L et al (2011) ISPyB: an information management system for synchrotron macromolecular crystallography. *Bioinformatics* 27:3186–3192
43. Kabsch W (2010) Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr D Biol Crystallogr* 66:133–144
44. Brehm W, Diederichs K (2014) Breaking the indexing ambiguity in serial crystallography. *Acta Crystallogr D Biol Crystallogr* 70: 101–109
45. Chirgadze NY, Kisselman G, Qiu W et al (2012) X-CHIP: an integrated platform for high-throughput protein crystallography. In: Benedict JB (ed) *Recent advances in crystallography*. InTech, Rijeka, pp 87–96
46. Dhoub K, Khan MC, Pflieger W et al (2009) Microfluidic chips for the crystallization of biomacromolecules by counter-diffusion and on-chip crystal X-ray analysis. *Lab Chip* 9: 1412–1421

47. Gavira JA, Toh D, Lopez-Jaramillo J et al (2002) Ab initio crystallographic structure determination of insulin from protein to electron density without crystal handling. *Acta Crystallogr D Biol Crystallogr* 58: 1147–1154
48. Gerdtz CJ, Elliott M, Lovell S et al (2008) The plug-based nanovolume microcapillary protein crystallization system (MPCS). *Acta Crystallogr D Biol Crystallogr* 64:1116–1122
49. Gerdtz CJ, Stahl GL, Napuli A et al (2010) Nanovolume optimization of protein crystal growth using the microcapillary protein crystallization system. *J Appl Crystallogr* 43: 1078–1083
50. Kisselman G, Qiu W, Romanov V et al (2011) X-CHIP: an integrated platform for high-throughput protein crystallization and on-the-chip X-ray diffraction data collection. *Acta Crystallogr D Biol Crystallogr* 67:533–539
51. May A, Fowler B, Frankel KA et al (2008) Diffraction-capable microfluidic crystallization chips for screening and structure determination. *Acta Crystallogr Sect A Found Adv* 64: C133–C134
52. Shim JU, Cristobal G, Link DR (2007) Using microfluidics to decouple nucleation and growth of protein crystals. *Cryst Growth Des* 7:2192–2194
53. Yadav MK, Gerdtz CJ, Sanishvili R et al (2005) In situ data collection and structure refinement from microcapillary protein crystallization. *J Appl Crystallogr* 38:900–905
54. Cipriani F, Rower M, Landret C et al (2012) CrystalDirect: a new method for automated crystal harvesting based on laser-induced photoablation of thin films. *Acta Crystallogr D Biol Crystallogr* 68:1393–1399
55. Mueller U, Darowski N, Fuchs MR et al (2012) Facilities for macromolecular crystallography at the Helmholtz-Zentrum Berlin. *J Synchrotron Radiat* 19:442–449
56. Hirata K, Kawano Y, Ueno G et al (2013) Achievement of protein micro-crystallography at SPring-8 beamline BL32XU. *J Phys Conf Ser* 425:012002

CD Spectroscopy: An Essential Tool for Quality Control of Protein Folding

Giuliano Siligardi and Rohanah Hussain

Abstract

The production of diffraction quality protein crystals for X-ray crystallography has been greatly accelerated by the development of high-throughput protein (HTP) methods, which enable a large number of crystallization conditions to be rapidly investigated. Monitoring sample quality and the effect of crystallization buffers on protein behavior in solution should be considered as part of the crystallization experiment. Circular Dichroism (CD) spectroscopy is the ideal technique for these tasks as it can be operated in a high-throughput mode. Using CD to screen ligand binding interactions could show whether protein function/activity is retained, altered, or lost under different crystallization conditions. In this chapter, several methods for high-throughput CD (HTCD) applied to the preparation of proteins for crystallization will be presented. Quality control (QC) of protein batches in terms of conformational folding is often disregarded in protein production. Examples of batch-to-batch variation in the local tertiary structure of aromatic side chain residues revealed by CD will be discussed. In some of the examples, the fact that ligand binding properties were affected by changes in folding clearly shows that the characterization of folding of recombinant protein batches should not be ignored but be implemented as an important part of protein quality control.

Key words Circular dichroism (CD), Synchrotron radiation circular dichroism (SRCD), Protein secondary structure, Local tertiary structure, High-throughput CD (HTCD), SRCD UV-denaturation assay

1 Introduction

In the last few decades, drug discovery has benefited from information about the three-dimensional structure of proteins at atomic resolution determined by X-ray crystallography [1] and NMR spectroscopy [2]. However, the number of resolved human protein structures deposited in the Protein Data Bank is still only a small fraction of the approximately 20,000 protein-coding genes in the human genome [3]. Proteins that cannot be studied by either X-ray crystallography due to their failure to crystallize or NMR due to size or irregular structure can be characterized by Circular Dichroism (CD) spectroscopy. CD is the differential absorption between left and right circularly polarized light of a chiral molecule [4] and therefore CD spectroscopy is sensitive to the absolute

configuration and conformation of chiral molecules. With the exception of glycine, amino acids are chiral molecules adopting either l or d stereoisomers which are nonsuperimposable mirror-imaged configurations. Natural proteins only consist of l amino acids.

Although of low resolution compared to X-ray and NMR methods, CD enables the rapid and direct characterization of protein folding in solution as a function of concentration, aqueous buffer composition, ionic strength, temperature, pH, detergents, chemical agents [4], and UV irradiation, age, and ligand binding interactions [5–7]. Importantly, CD spectroscopy can be used to assess the conformational behavior of a protein in a broad range of conditions to optimize protein formulation prior to carrying out high-resolution techniques such as X-ray crystallography and NMR spectroscopy. The production of good quality protein crystals for X-ray crystallography in terms of size and diffraction patterns has been greatly accelerated by high-throughput (HTP) screening of many crystallization conditions in multi-well plate formats. Combining this with high-throughput CD (HTCD) screening of protein formulations used for crystallization experiments provides added value. It can be used to characterize the sample in terms of conformational behavior in solution and may also identify structural features of the sample in a particular crystallization buffer that favor crystallization. Protein function, such as ligand binding interaction, can also be tested by HTCD to assess whether the protein function/activity is retained, altered, or lost under different crystallization conditions.

Another application of CD spectroscopy is monitoring batch-to-batch variation in recombinant protein production. This quality control (QC) is relevant not only to structural studies but also to the use of recombinant proteins for therapeutic purposes [8]. Recombinant proteins are often produced in several batches during the lifetime of a scientific project. The QC of protein batches is an important aspect of this process and HTCD offers a way of rapidly assessing protein folding and any variation between different batches of samples. For the biotechnology and pharmaceutical industry the QC of protein folding by CD spectroscopy could be used as a fingerprint of product quality.

In this chapter the application of SRCD using the Diamond B23 module A beamline to the quality control of proteins will be described.

2 Sample Preparation for CD Measurements

Scanning proteins in the far-UV region (185–260 nm) reveals details of secondary structure content, while measurements in the near-UV region (260–330 nm) reflect the local tertiary

conformation of aromatic side chain residues such as tryptophan, tyrosine, and phenylalanine and dihedral angles of the disulfide bonds [9]. The sample specification for SRCD measurements in the far-UV spectrum at Diamond B23 module A beamline is summarized as follows. The concentration of the protein sample needs to be 0.3–0.5 mg/ml in a total cell volume of 30–200 μ l for a 0.02 cm path length cell. The UV absorbance (A280 nm) of the solution should be between 0.5 and 1.5. Protein samples are often formulated with sodium chloride (NaCl) to maintain solubility. Since chloride anions are *not* optically transparent below 200 nm it is critical to keep the concentration of chloride ions (NaCl) and other buffer components (e.g., Tris-HCl, Hepes, MES) in the sample as low as possible, typically 10–20 mM. Higher salt concentration can be used provided the sample is at higher concentration allowing smaller path lengths of 2–50 μ m to be used according to Beer-Lambert's Law. Only 2-mercaptoethanol can be used as a reducing agent in the sample if this is required for stabilizing the protein. SRCD beamlines can penetrate by 5–10 nm the cutoff of the solvents/buffers encountered with benchtop instruments; however, any CD band observed below 170 nm should be treated suspiciously as due to noise artifact. For measurements in the near-UV region, the sample concentration is typically 1–2 mg/ml in total volume of approximately 500 μ l for a path length cell of 1 cm. Buffers containing 150 mM or higher NaCl can be used since the chloride anions are transparent in the near-UV region.

For a protein under a given environment condition, the content of protein secondary structure elements (α -helix, β -strand, β -turns, collagen type (PPII) and unordered) can be estimated quantitatively using the CONTIN, SELCON, and CDSSTR methods [10]. These methods can be applied using the suite of programs of bespoke benchtop CD instruments Chirscan (APL, UK), CD20 (Olis, USA), Jasco CD instruments (Japan), or accessing the following web sites: CD-Pro [<http://lamar.colostate.edu/~sreeram/CDPro/main.html>], DichroWed [<http://dichroweb.crysl.bbk.ac.uk/html/links.shtml>], and Diamond B23 CD-Apps [<http://www.diamond.ac.uk/Beamlines/Soft-Condensed-Matter/B23/manual/Beamline-software.html>, <http://confluence.diamond.ac.uk/display/B23Tech/CD+Apps+documentation>].

3 Instrumentation

Three screening systems are available for HTCD: the automated CD (ACD) Chirscan (Applied Photophysics, Leatherhead, UK [www.photophysics.com]) since 2012, the

high-throughput CD J1000 and J1500 Jasco spectropolarimeters [www.jasco.uk] since 2012, and HTCD-B23 module A beamline (Diamond Light Source, Chilton, Oxfordshire, UK [www.diamond.ac.uk]) since 2013. ACD, which is mainly used by Biotech and Pharma industries, uses a robot to inject the sample solution from 96-well or 384-well multiplates into a cuvette cell for CD spectral measurements. The automated cell cleaning, cell drying, sample injection, and CD measurement enable the processing of up to 200 samples per day [11]. The Jasco system that can accommodate two 96-well multiplates is based on a similar method of sample injection into a cuvette cell, CD measurement, waste collection, and cleaning cycle for the subsequent sample injection [www.jasco.de]. For both Chirascan and Jasco HTCD systems, the choice of the optimum concentration for the protein solutions to be used with the liquid transfer systems follows the same procedure described in Subheading 1, which is a protein concentration of 0.3–0.5 mg/ml for a 0.02 cm path length cell.

HTCD-B23 on the other hand is a multiplate CD technology that measures directly the synchrotron radiation CD (SRCD) from UV transparent fuse quartz 96-well or 384-well multiplates (Hellma, UK). The position of the transparent fuse quartz 96-well or 384-well multiplates containing the protein samples is controlled by a motorized X-Y stage in a vertical sample compartment (Fig. 1). The SRCD measurements are possible in this chamber due to the highly collimated micro light beam of about 1 mm (horizontal) \times 0.5 mm (vertical) of the B23 beamline that passing through the flatter central area of the meniscus solution is not affected by its curvature. The solution of the vertical chamber cannot be implemented with benchtop CD instruments because of their highly divergent and larger incident beam light (8 mm \times 10 mm), which will be affected by the meniscus curvature of the solution in the well and by the polarization artifacts produced by deflecting the incident light vertically. The important feature of this is that varying the volume of the sample solution in the well will vary linearly the solution height leading to reproducible measurements in the range from 1 (60 μ l for 96-well and 20 μ l for 384-well) to 8 mm equivalent path length (240 μ l for 96-well and 80 μ l for 384-well) (Fig. 1). Another important advantage of HTCD-B23 is that detergents, indispensable for membrane proteins production and viscous solutions, are not limiting or troublesome factors. The productivity of HTCD-B23 can be increased up to 300 samples per day. The use of the HTCD-B23 is open to UK scientific academia and industry with call for research proposals twice per year: in March and September [www.diamond.ac.uk].

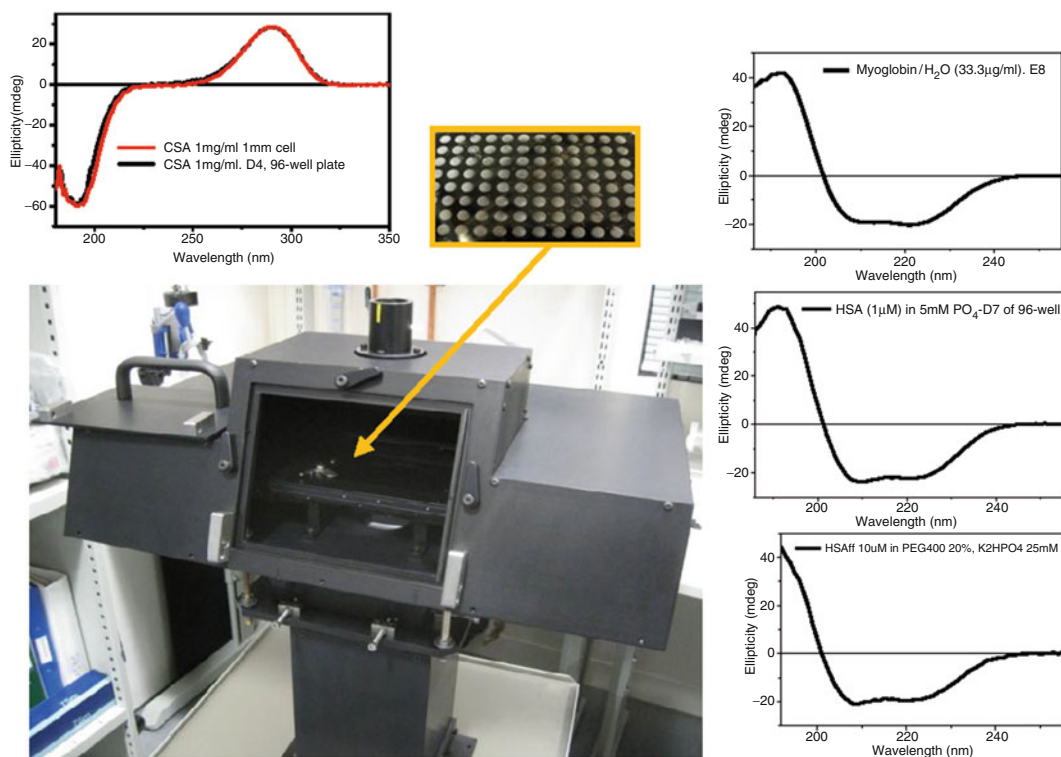


Fig. 1 Vertical sample compartment of Diamond B23 module A beamline. The chamber enables the SRCD measurements of horizontally positioned samples with respect to the incident monochromatic light. It has been designed to accommodate the 96- and 384-well multiplates made of fused quartz (Suprasil, Hellma). The *central insert* shows where the enlarged 96-well multiplate is located inside the chamber (*yellow arrow*). The *left insert* illustrates the SRCD spectrum of 60 μ l of 1 mg/ml of 1*S*-(+)-Camphor-10-sulfonic acid (CSA) measured in the D4 well (*black*) of the 96-well plate that was identical to that measured for the same solution in 1 mm pathlength cell (*red*). The *right inserts* illustrate from *top to bottom* the SRCD spectra of myoglobin in H₂O and human serum albumin (HSA) in 5 mM phosphate buffer and 20 % PEG 400 in 25 mM phosphate buffer respectively

4 Applications of SRCD

4.1 Protein Folding in Crystallization Buffers

The example in Fig. 2 illustrates the preliminary SRCD spectra of 96 equine skeletal muscle myoglobin (Sigma M0630) dissolved in MemGold2™ HT-96 crystallization screen (Molecular Dimensions (www.moleculardimensions.com)) (Table 1) using the 384-well multiplate. The MemGold2™ consists of 96 conditions of the most recent alpha helical membrane protein crystallization conditions that contain combinations of high ionic strength (100–200 mM) cations (Li, Na, K, Mg, Cd, Zn, and ammonium), anions (chloride, acetate, citrate, sulfate, cacodylate, formate), buffers (Tris, Hepes, Mes, Mms, Bis-Tris, Mada, Mops, Choline, Glycine, and phosphate) at various pHs from 4 to 9, and PEG of

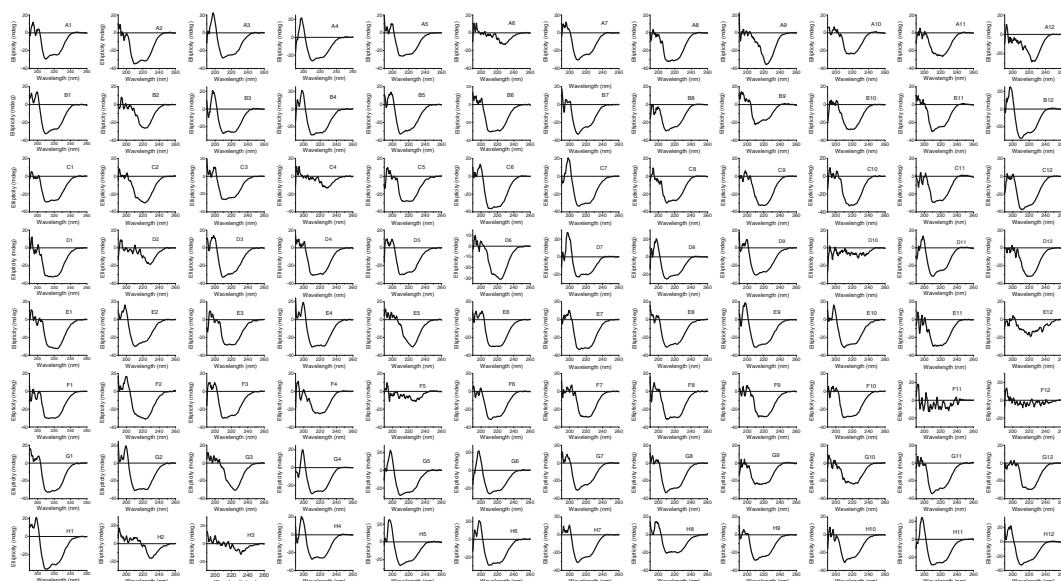


Fig. 2 SRCD spectra of 96 myoglobin solutions prepared from 96 crystallization buffer conditions of MemGold2™ solvent kit (see Table 1)

different molecular sizes (200–8,000) at high concentration from 9 to 66 % (Table 1). For each solution, 20 μ l of myoglobin (0.030 mg/ml) was transferred into 96 wells of the 384-well multiplate (Hellma, UK) used for this experiment. When possible, it is convenient to speed up the sample preparation by diluting a stock solution of high protein concentration with the appropriate volume of buffer to reach the final concentration of 0.010 mg/ml. Filling the well with 20 μ l of protein solution, the collected SRCD spectrum was identical to that measured with 1 mm path length cell (Hellma, UK). For each well, 20 μ l was found to be the minimum volume retaining spectral reproducibility in terms of path length. Using a 96-well multiplate instead would have required 60 μ l to achieve the same 1 mm path length due to the larger diameter of the well cell.

Figure 3 shows an example of data analysis in which the percentage of different secondary structure elements in the samples has been estimated from SRCD data using CONTIN/LL, a variant of CONTIN algorithm [10, 12, 13]. The estimation is calculated using a linear combination of CD spectra of reference soluble and membrane proteins with known secondary structure content assigned from X-ray coordinates [14]. For all CD methods, it is essential to be consistent with this method for assigning the protein secondary structure from X-ray atomic coordinates for the reference proteins.

Table 1
MemGold2™ Kit (Molecular Dimensions) for the screening of 96 crystallization conditions for alpha helical membrane proteins

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12
14 % PEG550 MME	44 % PEG 3000	10 % PEG 450	8 % PEG 1450 0.02 M CaCl ₂ 0.04 M MgSO ₄	32 % PEG 400 0.05 M NiSO ₄ 0.05 M LiCl	10 % PEG 400 3350 0.1 M	11.5 % PEG 4000	30 % PEG 400 0.1 M CdCl	20 % PEG 2000	31 % PEG 400	32 % PEG 400 0.2 M (NH ₄) ₃ PO ₄	14 % PEG4000
0.2 M MgCl ₂ , 0.005 M CdCl ₂ 0.1 M Tris	0.1 M KAc	0.08 M MgSO ₄	0.02 M MES	0.05 M Tris	0.1 M NaPO ₄ 0.1 M KPO ₄	0.1 M NaCl	0.1 M LiCl	0.2 M (NH ₄) ₂ SO ₄	0.2 M LiSO ₄	0.1 M (NH ₄) ₂ SO ₄ 0.1 M Na Citrate	0.05 M Na citrate
pH 7.5	0.01 M KCl	0.02 M NaCl 0.02 M MES	pH 6.5	pH 8.5	0.1 M Bis-Tris Propane	0.1 M LiSO ₄	0.1 M NaAc	0.1 M NaCl	0.1 M NaCl	pH 4.5	0.12 M KCl
	0.02 M Tris	pH 6			pH 7.5	0.1 M ADA	pH 4.5	0.1 M Na Citrate	0.1 M HEPES pH 7.0		0.08 M Bis-Tris
	pH 7				pH 6.5						pH 6.0
B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12
19 % PEG 1000	18 % PEG 2000 MME	3.5 % PEG 3350	14 % PEG 350 MME	35 % PEG 550 MME	28 % PEG 400	25 % PEG 350 MME	36 % MPD	11 % PEG 8000	26 % PEG 400	32 % PEG 400 3 % PEG 4000	
0.1 M NaCl	0.01 M NiSO ₄	0.02 M MgCl ₂	0.02 M NaCl	0.025 M MgCl ₂ 0.02 M MOPS	0.03 M MgCl ₂	0.04 M NaCl 0.04 M Tris	0.04 M MgAc 0.1 M MES	0.05 M ZnAc	0.05 M MgAc	0.05 M MgAc	0.066 M NaCl
0.15 M (NH ₄) ₂ SO ₄	0.1 M NaCitrate pH 6.0	0.02 M MES	0.05 M ME S pH 5.5	pH 7.0	0.1 M MES	pH 8	pH 6	0.05 M ADA	0.1 M MES	0.1 M Glycine pH 9.5	0.02 M Tris
0.01 M MES		pH 6.0			pH 6.5			pH 6.3	pH 6.5		pH 7.5

(continued)

Table 1
(continued)

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12
pH 6.5											
C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
30 % PEG 2000 MME	14 % PEG5000 MME	2.0 M (NH ₄) ₂ SO ₄	22 % PEG250 DME	13 % PEG 8000	17 % PEG 3350	18 % PEG 4000	18 % PEG 200	22 % PEG 4000	22 % PEG 8000	23 % PEG 3350	32 % PEG 400
0.075 M MgCl ₂	0.08 M MgAc	0.05 M ZnAc	0.087 M (NH ₄) ₂ SO ₄	0.1 MgCl ₂	0.1 M Mg formate	0.1 M KCl	0.1 M KCl	0.1 M MgAc	0.1 M CaAc	0.1 M (NH ₄) ₂ SO ₄	0.1 M KCl
0.1 M Na Cacodylate	0.1 M Na citrate pH 6.0	0.1 M MES	0.5 M Tris	0.1 M Tris	0.1 M MOPS	0.1 M Bis- Tris	0.1 M KPO ₄	0.1 M MES	0.1 M MES	0.1 M HEPES	0.1 M MES
pH 6.5		pH 6	pH 7.0	pH 7.5	pH 7	pH 6	pH 7.5	pH 6.0	pH 6	pH 8.5	pH 6
D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
36 % PEG 300	45 % PEG 550 MME	35 % PEG 400	13 % PEG 4000	14 % PEG 2000 MME	19 % PEG 6000	19 % PEG 6000	19 % PEG 3350	20 % PEG 350 MME	20 % PEG 3350	33 % PEG 1000	24 % PEG 400
0.1 M NaCl	0.1 M NaCl	0.15 M CaCl ₂ 0.1 M Glycine pH 9.0	0.2 M (NH ₄) ₂ SO ₄	0.2 M choline Cl	0.2 M NaCl	0.05 M NaCl	0.2 M Mg formate 0.05 M Tris	0.2 M CaCl ₂	0.2 M (NH ₄) ₂ NO ₃	0.5 M MgCl ₂	0.2 M CaAc
0.1 M MES	0.1 M Bicine		0.5 M ADA	0.1 M Tris	0.05 M MOPS	0.05 M MOPS	pH 8	0.1 M MES	0.05 M HEPES	0.02 M LiCl	0.1 M HEPES
pH 6.5	pH 9		pH 6.5	pH 7.5	pH 7.0	pH 7.0		pH 5	pH 7.0	0.02 M Glycine	pH 7
E1	E2	E3	E4	E5	E7	E7	E8	E9	E10	E11	E12

28 % PEG 400	29 % PEG 400	29 % PEG 400	31 % pentaerythritol ethoxylate	35 % PEG 3350	38 % PEG 400 0.2 calcium acetate—none 0.1 Sodium Acetate 5.0	38 % PEG 400 0.2 sodium chloride—none 0.1 MOPS 7.5	400 2.0 M Ammonium Sulfate 0.2 sodium chloride—none 0.1 Sodium Cacodylate 6.5	12 % PEG 4000 0.225 ammonium sulfate—none 0.05 Sodium Acetate 4	33 % PEG 400	22 % PEG 3000 0.25 M Mg formate	40 % PEG 1000 0.25 M MgCl ₂
0.2 M NaAc	0.2 M NaCl	0.2 M NaCl	15/04 0.2 ammonium formate—none 0.1 Tris 7	0.2 M (NH ₄) ₂ SO ₄ —none 0.1 Tris 8.5					0.23 M NaCl	0.1 Na Cacodylate	0.1 M Tris
0.1 M MES	0.05 M CaAc	0.1 M HEPES							0.05 M NaAc	pH 6.5	pH 8.5
pH 6.5	pH 5.0	pH 7							pH 4.5		
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
25 % PEG 400	33 % PEG 550 MME	34 % PEG 400	14 % PEG 4000 0.32 M LiCl	12 % PEG 4000 0.34 M (NH ₄) ₂ SO ₄	11 % PEG 600	22 % PEG 400	10 % PEG 3350	32 % PEG 400	12 % PEG 400	15 % PEG 4000	16 % PEG 4000
0.3 M LiSO ₄	0.3 M NH ₃ formate	0.3 M BaCl ₂	0.1 M Na Citrate	0.1 M Na citrate pH 5.5	0.35 M LiSO ₄	0.37 M KNO ₃	0.4 M (NH ₄) ₂ SO ₄	0.05 M NaCl	0.4 M KCl	0.4 M ammonium thiocyanate	0.4 M Na thiocyanate
0.1 M MES	0.05 M Tris	0.1 M MES	pH 5.5		0.1 M NaAc	0.1 M MES	0.1 M MES	0.04 M MgCl ₂	0.05 M HEPES pH 7.50	0.1 M NaAc	0.1 M NaAc
pH 6.5	pH 9.0	pH 6			pH 4.5	pH 6.5	pH 6.5	0.1 M HEPES		pH 4.5	pH 4
								pH 7.5			
G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12

(continued)

Table 1
(continued)

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12
20 % PEG 400	21 % PEG 350 MME	11 % PEG 4000	9 % PEG 8000	11 % PEG 20000	13 % PEG 400	14 % PEG 6000	17 % PEG 350 MME	22 % PEG 350	24 % PEG 300 400	24 % PEG 1500	28 % PEG 600
0.5 M KCL	0.5 M MgCl ₂	0.8 M K formate	0.1 M MOPS	0.1 M MES	0.1 M MES	0.1 M MES	0.05 M Tris	0.07 M Na citrate pH 4.5	0.05 M ADA	0.1 Na Cacodylate pH 6.5	0.1 M HEPES
0.05 M HEPES	0.05 M Tris	0.1 M NaAc	pH 7	pH 6.0	pH 6.5	pH 5.5	pH 7.5		pH 6.5		pH 7.5
pH 6.5	pH 7.5	pH 5.0									
H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12
28 % PEG 400	30 % PEG 400	31 % PEG 600	32 % PEG 550 MME	33 % PEG 400	34 % PEG 3350	44 % PEG	200 65 % MPD	2.75 M NH ₄ Cl	2.8 M NH ₄ Cl	3.0 M (NH ₄) ₂ SO ₄ 0.1 M MES	3.25 M 1,6- Hexanediol
0.05 M Tris	0.1 M Bicine	0.1 M ADA	0.10 M Tris	0.1 M HEPES	0.18 M Na citrate pH 4.0	0.1 M Na	0.1 M Tris	0.0 25 M Bis-Tris pH 7	0.075 M HEPES	pH 5.5	0. 01 M HEPES
pH 8.5	pH 9	pH 7.0	pH 8.5	pH 7.5		pH 8.5	pH 8		pH 7.5		pH 7.5

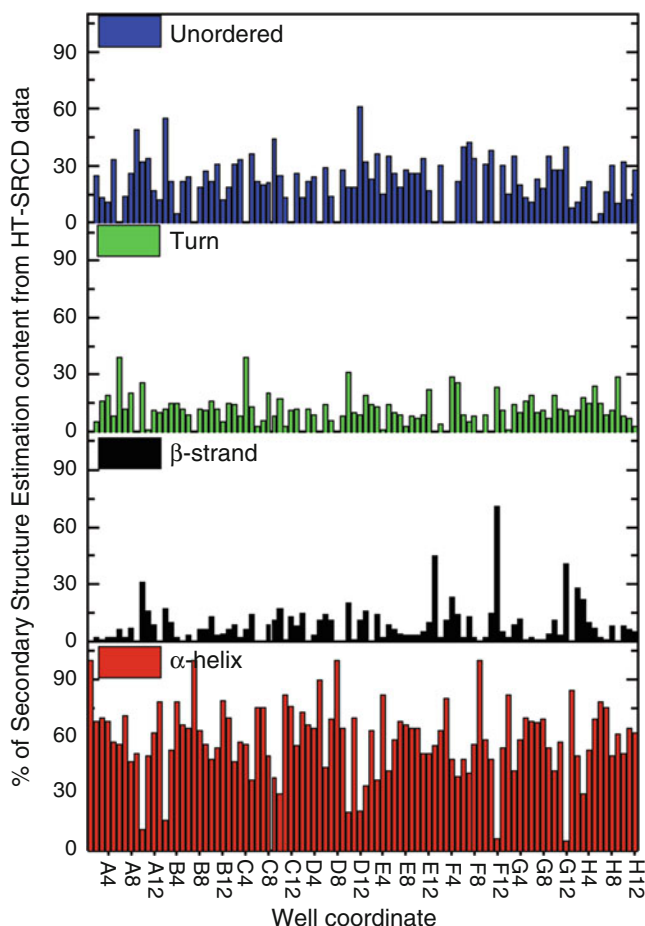


Fig. 3 Bar chart of the percentage of Protein Secondary Structure Estimation (PSSE) for unordered (*blue*), turn (*green*), β -strand (*black*), and α -helix (*red*) conformations from 96 SRCD data of myoglobin in MemGold2™ crystallization solvent conditions (see Fig. 2) using CONTIN/LL [5]

In some of the wells (A6, A10, B2, C4, D2, D10, E12, F5, F11, F12, H2, and H3 of Fig. 2), the high salt content interfered with the measurement leading to an underestimate of the alpha helical content. Spectral CD changes in the far-UV region associated with protein folding were observed in several wells (Fig. 3). The protein secondary structure content estimated from CD data indicated the appearance of significant β -sheet content for myoglobin under several formulation conditions (Fig. 3).

The Peltier 6-cell turret holder, introduced by On-Line Instrument Systems [www.olisweb.com] since 2005, available at B23 since 2011 and from 2013 for Chirascan and Jasco CD instruments [www.jasco.uk], can be used to carry out experiments as a function of many variables such as temperature, pH, and ligand interaction. The 6-cell Peltier turret on B23 (Fig. 4) allows the

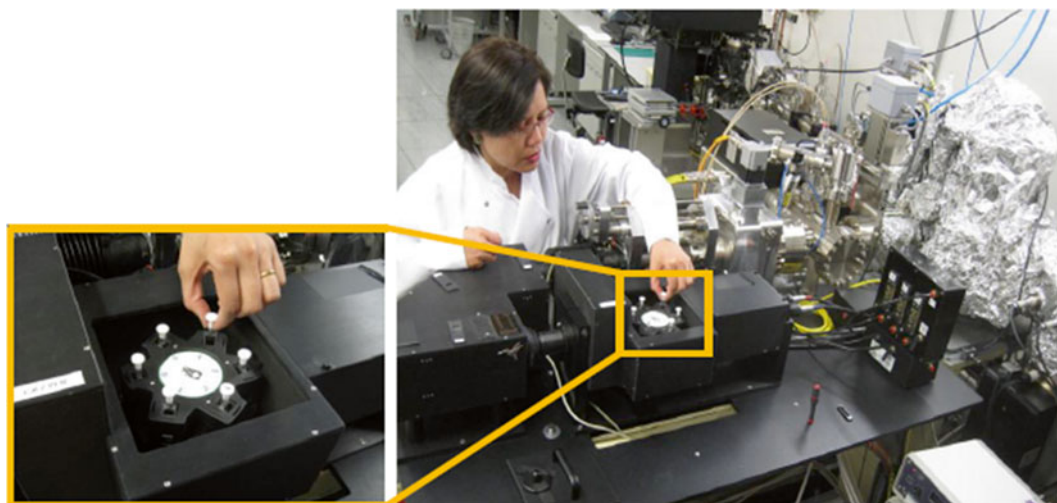


Fig. 4 6-Cell Turret of Diamond B23 module B beamline. The turret is used for the SRCD UV-protein denaturation assay experiment in the far-UV region and/or variable temperature measurements in the 5–95 °C temperature range

measurements of six rectangular cuvette cells from 0.1 to 10 mm path length [Hellma (www.hellma-analytics.com) and Starna (www.starnacells.com)] as well as demountable from 0.01 to 0.1 mm path length [Hellma]. The measurements can be conducted on the same sample as a function of solvent, buffer, pH, concentration, ligand binding interaction, and temperature. One of the important applications of B23 beamline, the protein UV-denaturation assay [5–7], makes use of the 6-cell turret. In Fig. 5a are illustrated the conformational behaviors of a monoclonal antibody (Mab1) under six different formulation conditions when irradiated in the far-UV region under the same parameters of photon flux, irradiation time number of repeated consecutive scans, and antibody concentration. Each sample was irradiated by scanning 30 consecutive repeated spectra in the 180–260 nm region corresponding to an irradiation time of approximately 90 min for that wavelength region. The collapse of the positive CD band at about 200 nm associated with the π - π^* transition of the β -sheet conformation is a direct indication of the loss of secondary structure as the antibody unfolds and can be related to a reduction in protein stability. The complete experiment was carried out overnight over 9 h as a single multiscrypt experiment that controlled the operation of the B23 module B beamline. The rate of UV-irradiation assay revealed that Mab1 in buffer formulation EC4 was the most stable while Mab1 in EC6 the least stable (Fig. 5b). In terms of relative stability, the six formulations can be ranked qualitatively as follows: EC4 > EC5 > EC1 = EC2 \gg EC6.

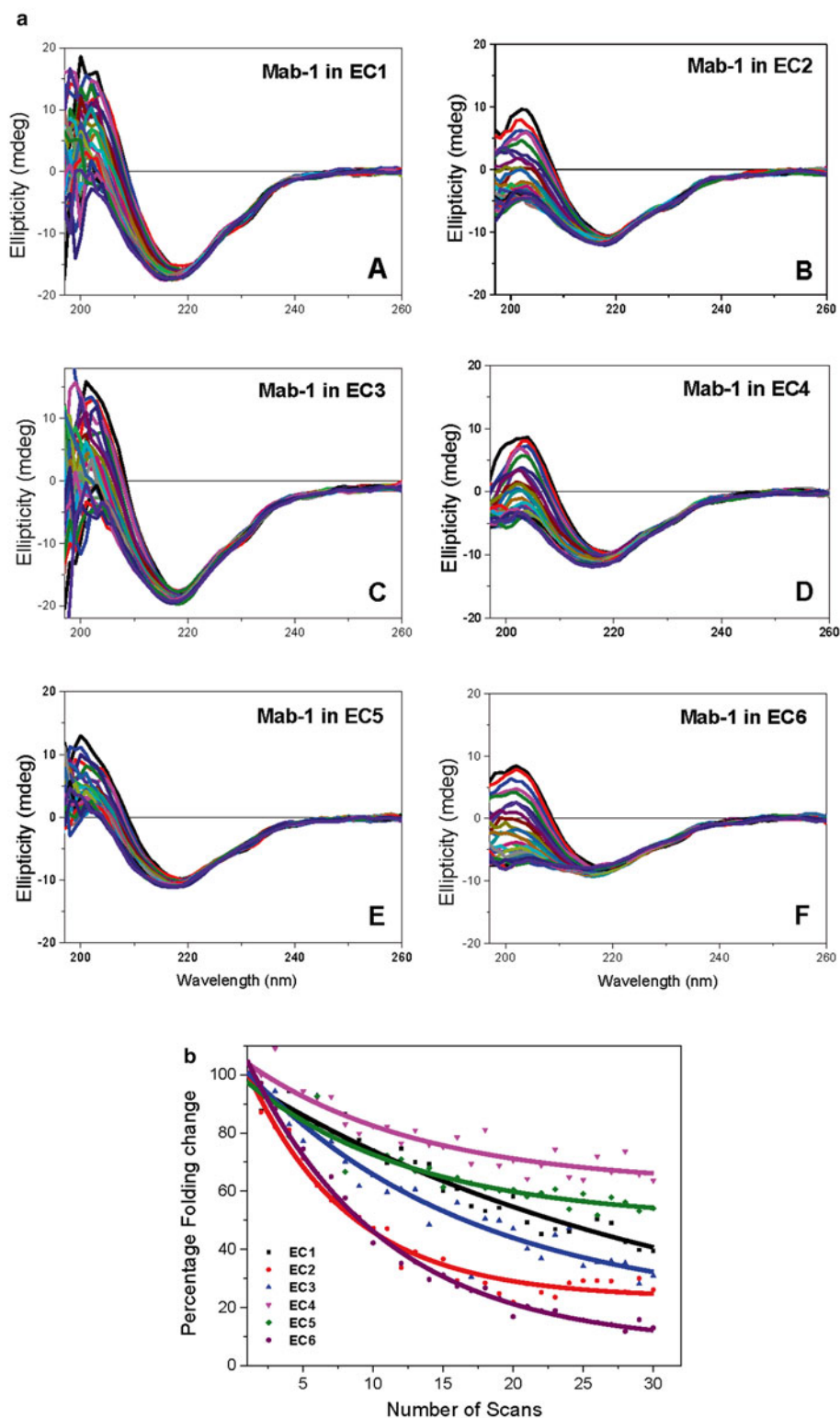


Fig. 5 (a) SRCD UV-denaturation assay in the far-UV region of a monoclonal antibody (Mab1) in six different formulations (EC1 to EC6). In this example the assay consisted of 30 repeated consecutive scans measured with Diamond B23 beamline. (b) Rates of UV-denaturation for Mab1 in six different formulations. From the rate of UV-denaturation, Mab1 in EC6 is the least stable while Mab1 in EC4 is the more stable one

4.2 Quality Control of Protein Folding and Determination of Ligand Binding by SRCD

The study of the function/activity of recombinant proteins in solution includes the analysis of folding that is the secondary, tertiary, and quaternary structure, determination of ligand binding all of which can be measured by SRCD. Critically, if several different batches of a protein are studied, it is important to establish that protein folding is the same for each batch. Subtle changes in the purification procedure can, for instance, produce protein batches of different quality and one effect that is commonly observed is a change in protein folding. CD spectroscopy is the ideal technique to characterize and estimate directly the protein folding in terms of secondary structure content and in particular the conformational changes as a function of environmental perturbations such as solvent composition, ionic strength, concentration, pH, temperature, chemical and detergent agents, and ligand interaction.

Ligand binding interaction can be determined both qualitatively and quantitatively with the method developed by Siligardi used to calculate dissociation constant (K_d) following [15–17]. In this method, the data of the CD titrations are reported as difference CD spectra calculated by subtracting from the spectra of the protein–ligand mixtures the equivalent spectra for each addition of the chiral ligand. In this way the spectral changes observed in the difference CD spectra are unambiguously indicative of ligand binding. The dissociation constant K_d is determined from the CD data at fixed wavelength as a function of ligand concentration using a nonlinear regression analysis [15]. Many other techniques, such as fluorescence, SPR, ITC, and AUC, are used to determine protein ligand binding interactions; however, only CD spectroscopy determine directly whether the interaction has induced changes in protein secondary and/or tertiary structures.

Reproducible gel filtration and mass spectrometry data for different batches of a recombinant protein do not necessarily mean that folding of the proteins is identical. Therefore, the implementing a quality control (QC) step for folding is therefore a must. The following examples illustrate some of the common problems that are often encountered in studying recombinant proteins in solution that have been revealed by CD spectroscopy.

4.2.1 Example 1: Yeast Heat Shock Protein 90 (hsp90)

Heat shock protein 90 (hsp90) is a ubiquitous and abundant molecular chaperone that mediates protein folding and activation of many signal transduction and cell regulatory proteins. Figure 6 shows the CD spectra of nine recombinant hsp90 batches produced over 3 years (from January 1998 till February 2001) during collaborative research with Dr. Christopher Prodromou and Professor Laurence Pearl [18]. The CD spectra recorded in the near-UV region were characteristic of the local tertiary structure of the side chains of aromatic amino acid residues (Trp, Tyr, and Phe) [19] and dihedral angle of disulfide bonds [5]. Out of the nine batches, two batches (09/02/2000 and 22/02/2000 of Fig. 6) showed significant CD spectral changes in the near-UV region at

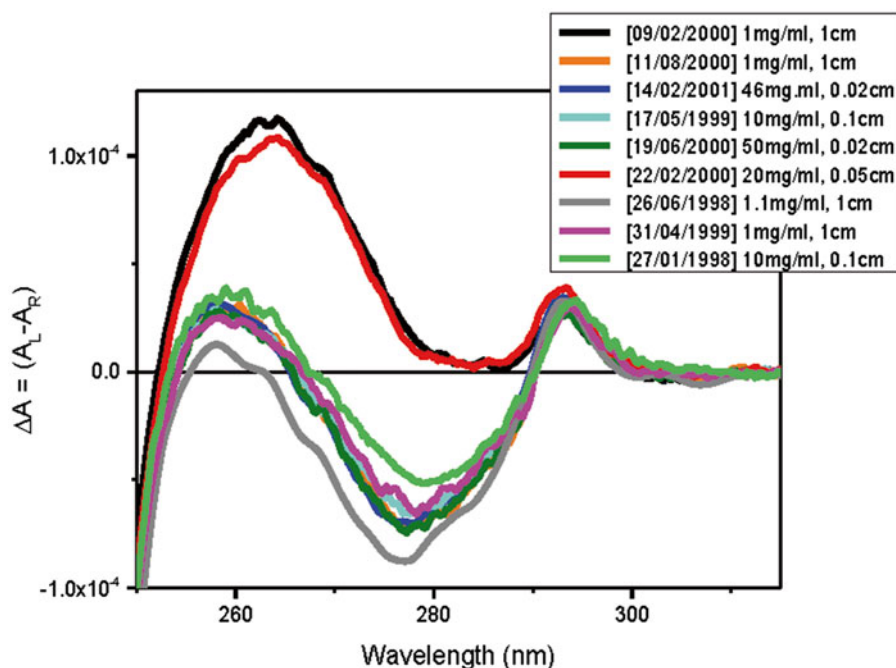


Fig. 6 CD spectra in the near-UV region of nine batches of recombinant Hsp90 protein measured within 4 years period from 1998 to 2001. The spectra were measured with Jasco J720 CD spectropolarimeter

around 280 nm associated with differences in local tertiary structure of Tyr side chain residues compared to the other similar six batches. However, the protein binding properties of hsp90 were not affected as the two batches representing the two types of local tertiary structure (batches 22/02/2000 and 19/06/2000 of Fig. 6) still bound to hP50 protein with the same dissociation constant K_d of 250 nM (Fig. 7). In addition, there was difference in the ATPase activity of the two batches of hsp90. One interesting observation was that for batch 22/02/2000 (Fig. 6) the complex remained soluble throughout the titration with hP50 while for the other hsp90 batch the complex started to precipitate at higher hP50 molar ratio than 1:1 [15].

4.2.2 Example 2: Link Domain of TSG-6 Glycoprotein

The Link module of human TSG-6 glycoprotein is involved in the formation of the extracellular matrix and cell migration by interacting with hyaluronan 10 (HA_{10}) [20]. This interaction was studied using two different batches of the Link module protein (batch 1 and 2) by CD spectroscopy. The CD spectral profiles of the two protein batches were not superimposable in the near-UV region (Fig. 8) indicating significant differences in the local structure of the side chains of aromatic residues. The Link module glycoprotein contains two tryptophan (Trp), eight tyrosine (Tyr), and three phenylalanine residues (Phe) (1tsg.pbd). The positive CD bands of the two batches

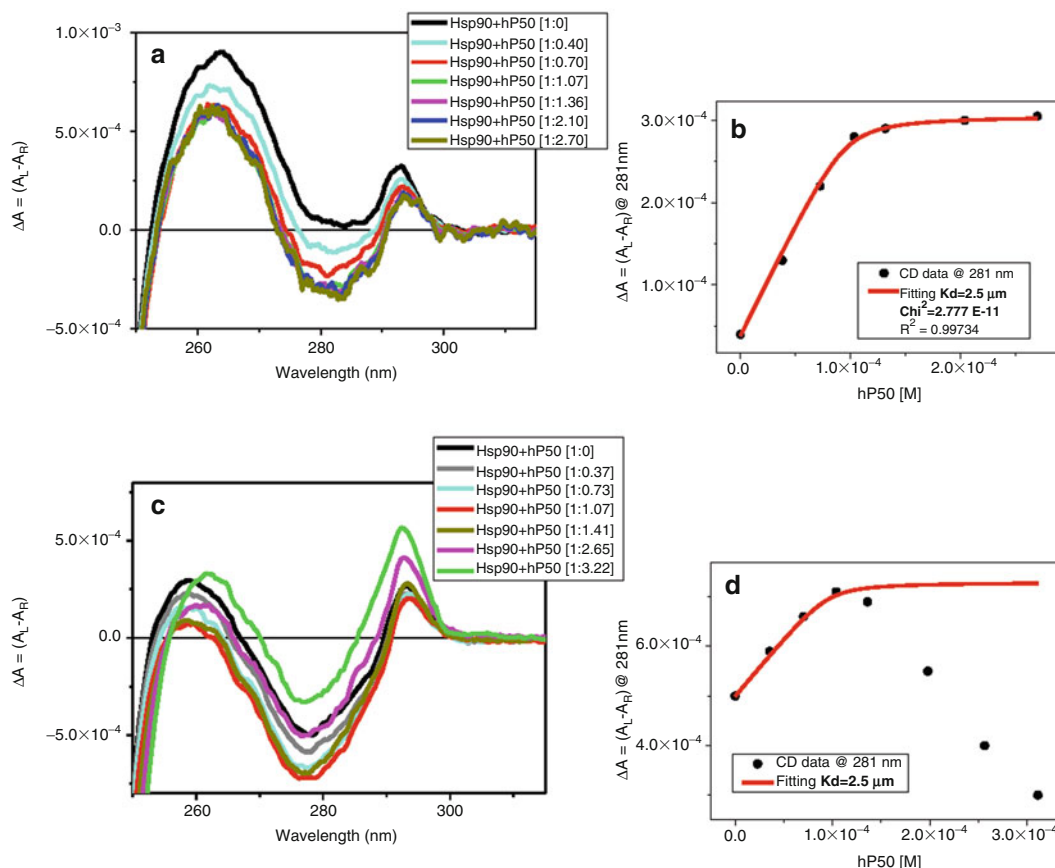


Fig. 7 (a) Difference CD spectra of the titration of hP50 into Hsp90 (batch 22/02/2000). The difference spectra were calculated subtracting from the each spectrum of (Hsp90 + hP50) mixtures at [1: n] molar ratio the equivalent spectrum of hP50 at [n] molar ratio. (b) Plot of difference CD intensity reported in $\Delta A = (A_L - A_R)$ measured at 281 nm versus the ligand hP50 concentration. The best fitting of the experimental data of (a) calculated using the nonlinear regression method of Siligardi et al. [15] was achieved for a $K_d = 0.250 \mu\text{M}$ and $R^2 = 0.997$. (c) Difference CD spectra of the titration of hP50 into Hsp90 (batch 19/06/2000). (d) Plot of difference CD intensity reported in $\Delta A = (A_L - A_R)$ measured at 281 nm versus the ligand hP50 concentration. The best fitting of the experimental data of figure (c) calculated using the nonlinear regression method of figure (b) was achieved for a $K_d = 0.250 \mu\text{M}$. Due to protein association [15], only the first four experimental data were fitted and therefore R^2 could not be calculated

at around 298 nm, assigned to the Trp residues, indicated a similar local tertiary structure (Fig. 8a), whereas a major CD spectral difference below 290 nm, assigned to the Tyr residues, indicated a significantly different local tertiary structure (Fig. 8a). The affect on hyaluronan binding was investigated by CD spectroscopy.

For the Link protein batch 1, the addition of (HA)₁₀ at molar ratio of [1:1] induced an increase in the intensity of the 285 nm positive CD band assigned to a Tyr residue that reached saturation at 2:1 molar ratio (Fig. 8b). The fact that the rest of the CD spectrum remained unchanged indicated that the other aromatic residues in the protein the two Trp residues and the majority of Tyr

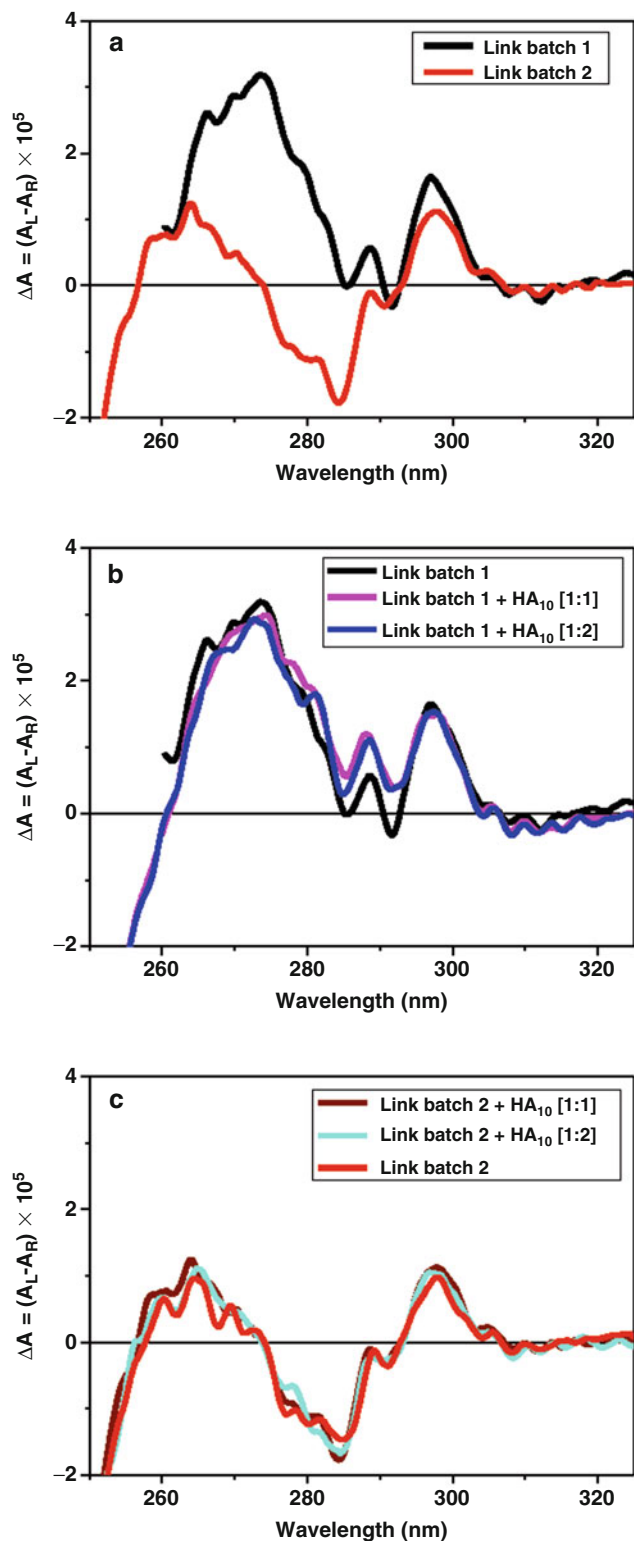


Fig. 8 (a) Near-UV CD spectra of batch 1 (*black*) and 2 (*red*) of TSG-6 Link Module protein. (b) CD spectra of bath 1 of TSG-6 Link module (*black*) with 1 (*pink*) and 2 (*blue*) molar ratios of hyaluronan 10 (HA₁₀). (c) CD spectra of bath 2 of TSG-6 Link module (*red*) with 1 (*brown*) and 2 (*cyan*) molar ratios of hyaluronan 10 (HA₁₀). The CD spectra were measured with Jasco J720 spectropolarimeter

residues were not involved in the ligand binding interactions (Fig. 8b). This was consistent with the NMR data of the complex formation [21]. For the Link protein batch 2, the hyaluronan binding interaction study showed no detectable CD changes in the near-UV region (Fig. 8c). This does not rule out some interaction between the Link protein batch 2 and hyaluronan as in the far-UV region, the CD spectra of Link batch 2 with and without hyaluronan were not superimposable indicating ligand binding interaction (data not shown). Thus the CD measurements in the far- and near-UV regions confirmed the hyaluronan binding property of the Link module protein but it also revealed folding differences between the two samples of the protein.

4.2.3 Example 3: K⁺ Transporter TrkA Peripheral Membrane Protein

TrkA is a peripheral membrane protein of the Trk system that requires ATP for the transport of K⁺ in prokaryotic and eukaryotic cells [22]. Five samples (p001, p002, p003, p005, and p006) of purified *H. influenzae* TrkA peripheral membrane protein in 500 mM NaCl, 50 mM Tris pH 7.5, 5 % glycerol, and 0.1 mM TCEP solutions were investigated using CD spectroscopy. The protein does contain, nine tyrosine residues and but no tryptophans or disulfide bonds [23]. In the near-UV region (250–315 nm) the CD spectra of the five equimolar solutions of TrkA were not identical (Fig. 9) indicating local conformational differences in the

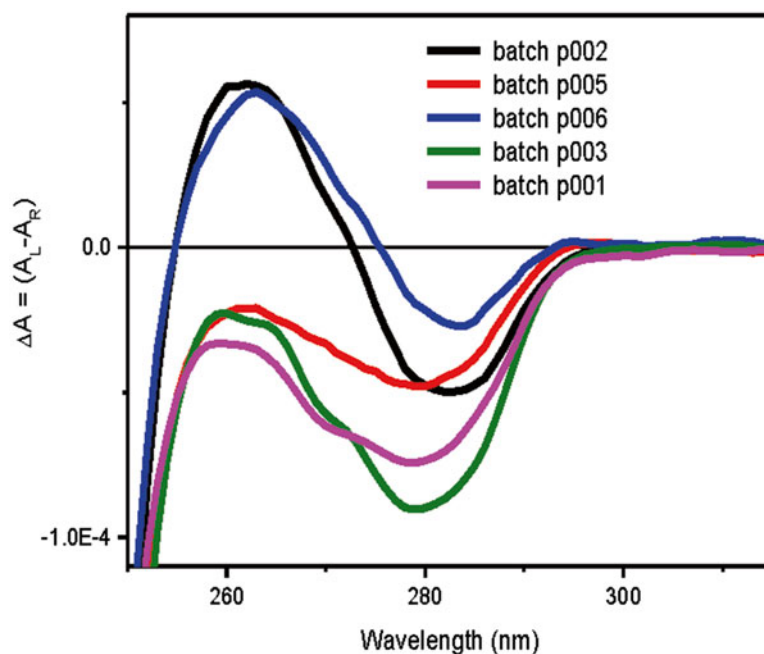


Fig. 9 Near-UV CD spectra of five batches (batch p001 (pink), p002 (black), p003 (green), p005 (red), and p006 (blue)) of 20 μ M TrkA peripheral membrane protein in 500 mM NaCl, 0.5 % glycerol, 0.1 mM TCEPT, and 50 mM Tris measured in 1 cm pathlength cell using Chirascan Plus CD spectropolarimeter (Applied Photophysics Ltd)

aromatic side chains of the Tyr residues. To investigate the effect of these conformational differences on the function of TrkA protein, a CD titration with ATP ligand was conducted for each protein batch. ATP was found to bind to TrkA and two main behaviors were observed. In Fig. 10, the data for the difference CD spectra, calculated by subtracting from each mixture of [TrkA+ ATP] at different (1:*n*) molar ratios the equivalent spectra for ATP (*n*), were fitted using a nonlinear regression method [15] (inserts of Fig. 10). The results for Sample p003 indicated that the TrkA has a single ATP binding site whereas the data for samples p001 and p006 (Fig. 10) suggested that there are two ATP binding sites. The results for the other batches of TrkA fitted with either a one-site or two-site model (Fig. 11). From the local tertiary structure, the batches of TrkA protein could be clustered into two groups: p002 with p006 and p001 with p003 and p005 as having qualitatively similar CD spectra (Fig. 9) while according to ATP binding stoichiometry, the groups could be rearranged differently: p002 with p003 and p001 with p005 and p006 (Fig. 11). Further studies will be required to clarify the functional significance of these observations but these behaviors would not have been fully revealed without the analysis of different batches of protein by CD spectroscopy.

5 Summary

HTCD offers a rapid way of assessing protein folding in solution and the effect of buffer conditions on secondary structural features. This in turn may inform how a protein sample behaves in crystallization trials. HTCD also allows the screening of the binding properties of the proteins in solutions under different conditions including in crystallization buffers. Another important application of HTCD is monitoring variations in protein folding between different batches of the same protein. In fact, the quality control of conformation and binding interactions of recombinant proteins by CD spectroscopy ought to play a more important role in structural biology.

Acknowledgements

We would like to thank Dr. Christopher Prodromou and Prof Laurence Pearl for providing the batches of Hsp90 proteins, Prof Antony Day for TSG-6 Link Module protein, and Dr. Petra Lukacik and Dr. Martin Walsh for TrkA protein. We would like to thank Dr. Tamas Javorfi for his assistance in commissioning the vertical chamber for Diamond B23 module A beamline.

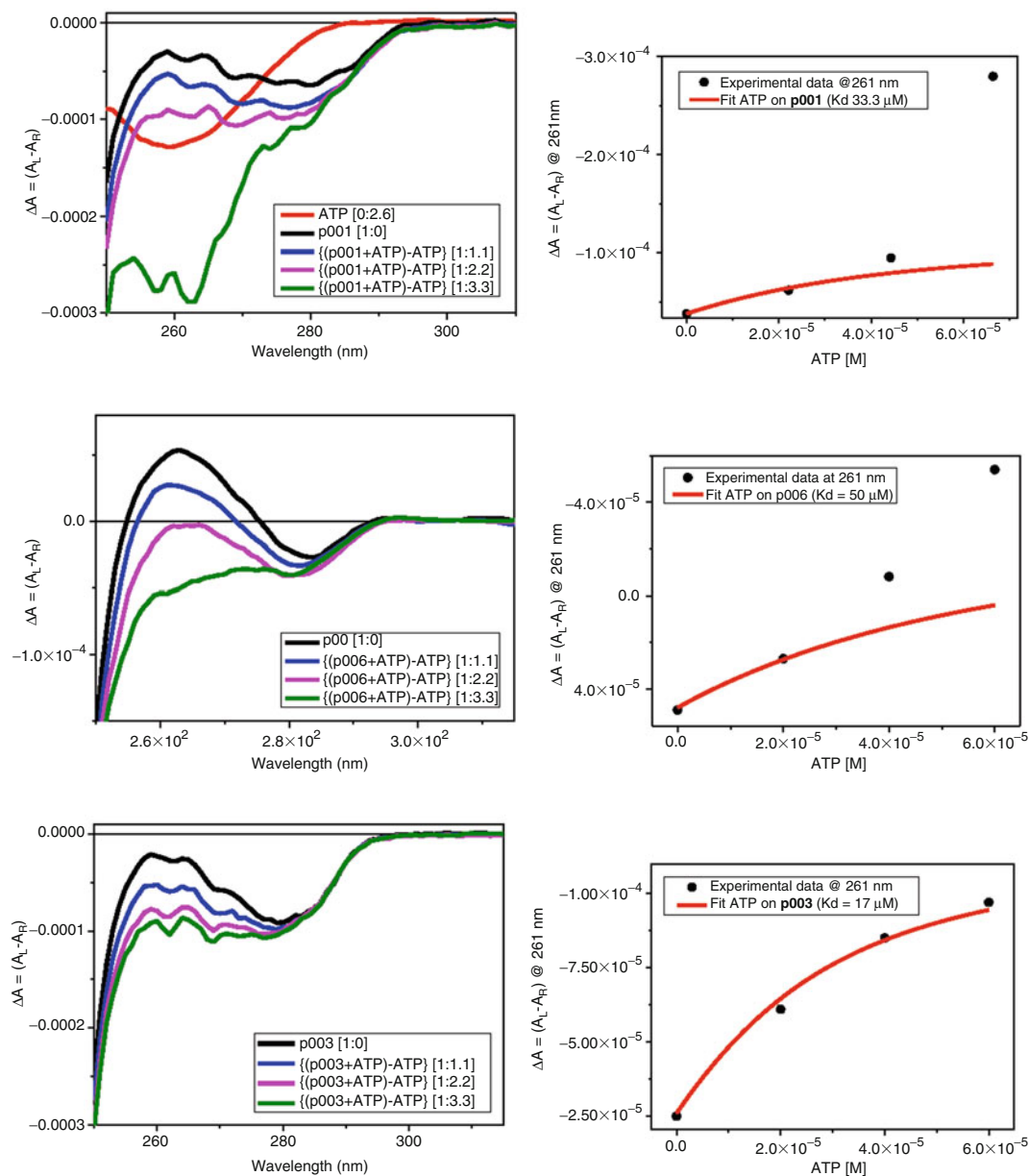


Fig. 10 (Top left) Difference CD spectra of the titration of ATP into TrkA (batch p001) in a 1 cm pathlength cell. The difference spectra were calculated subtracting from the each spectrum of (TrkA+ ATP) mixtures at [1:*n*] molar ratio the equivalent spectrum of ATP at [*n*] molar ratio. (Top right) Plot of difference CD intensity reported in $\Delta A = (A_L - A_R)$ measured at 261 nm versus the ligand ATP concentration. The best fitting of the experimental data calculated using the nonlinear regression method of Siligardi et al. [15] was achieved for the first two data supportive of ATP stoichiometry of 2. (Middle left) Difference CD spectra of the titration of ATP into TrkA (batch p006). (Middle right) Plot of difference CD intensity reported in $\Delta A = (A_L - A_R)$ measured at 261 nm versus ATP concentration. The best fitting of the experimental data was achieved like for batch p001 for the first two data supportive of ATP stoichiometry of 2. (Bottom left) Difference CD spectra of the titration of ATP into TrkA (batch p003). (Bottom right) Plot of difference CD intensity reported in $\Delta A = (A_L - A_R)$ measured at 261 nm versus ATP concentration. The best fitting of all experimental data was achieved indicating an ATP stoichiometry of 1

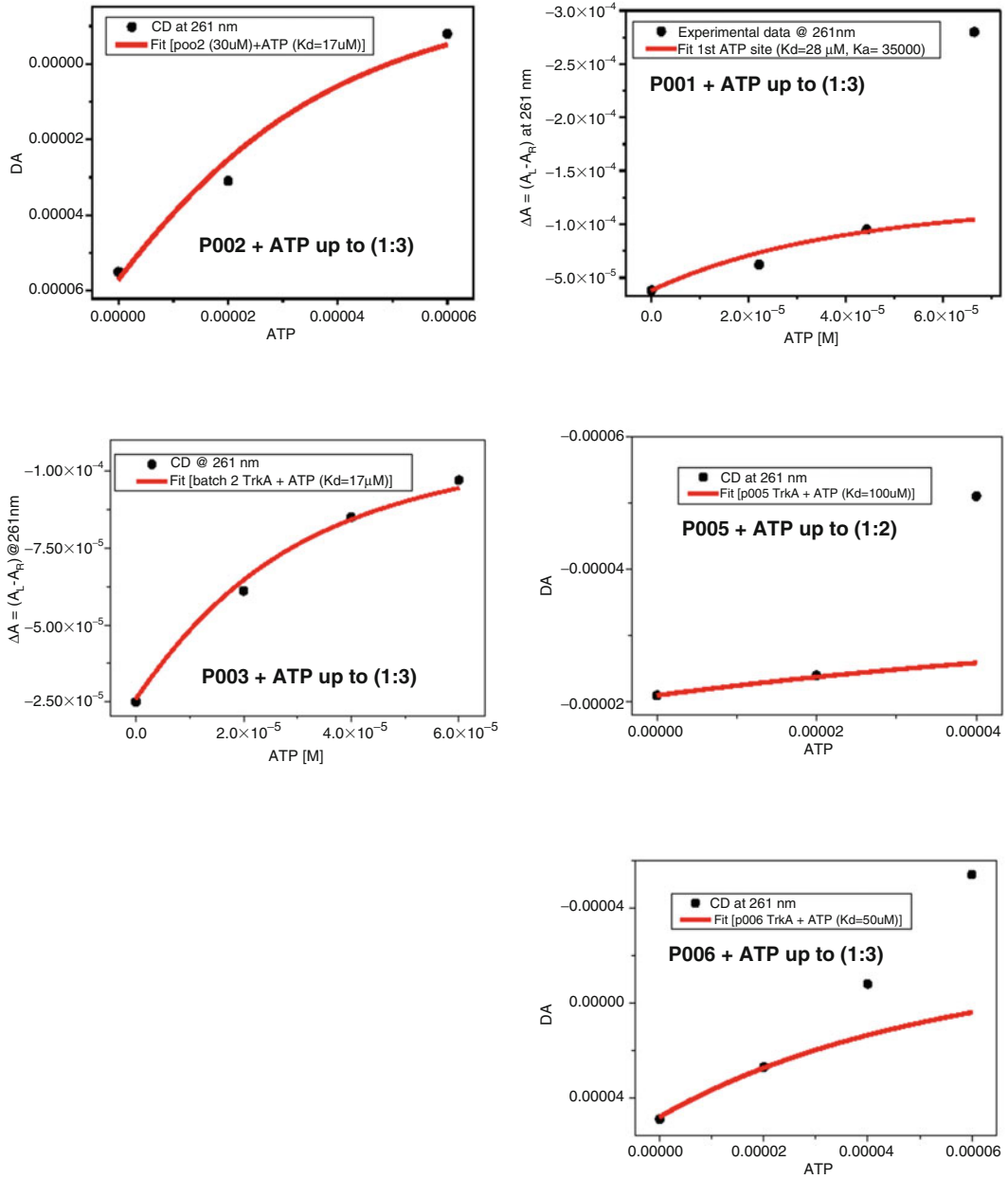


Fig. 11 Fitting of the experimental data of the plot CD data in ΔA versus ATP concentration for all five TrkA batches. The fittings of the experimental data measured with Chirascan Plus were calculated using the non-linear regression method of Siligardi et al. [15]

References

1. Carvalho AL, Trincão J, Romão MJ (2009) X-ray crystallography in drug discovery. *Methods Mol Biol* 572:31–56
2. Pellecchia M, Bertini I, Cowburn D et al (2008) Perspectives on NMR in drug discovery: a technique comes of age. *Nat Rev Drug Discov* 7:738–745
3. Pennisi E (2012) ENCODE project writes eulogy for junk DNA. *Science* 337:1159–1160
4. Fasman GD (1996) Circular dichroism and the conformational analysis of biomolecules. Plenum, New York
5. Hussain R, Jávorfí T, and Siligardi G (2012) Spectroscopic Analysis: Synchrotron Radiation Circular Dichroism. In: Carreira E.M. and Yamamoto H. (eds.) *Comprehensive Chirality*, Volume 8, pp. 438–448. Amsterdam: Elsevier
6. Hussain R, Jávorfí T, Siligardi G (2012) Circular dichroism beamline B23 at the Diamond light source. *J Synchrotron Radiat* 19:132–135
7. Calzolari L, Laera S, Ceccone G et al (2013) Gold nanoparticles' blocking effect on UV-induced damage to human serum albumin. *J Nanoparticle Res* 15:1412–1416
8. ICH, Topic Q1B, EMEA (1998) CPMP/ICH/279/95
9. Siligardi G, Campbell MM, Gibbons WA, Drake AF (1991) Conformational analysis of the melanin concentrating hormone (MCH) by CD spectroscopy: disulphide bridge and aromatic tyrosyl contributions. *Eur J Biochem* 206:23–29
10. Sreerama N, Woody RW (2000) Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON and CDSSTR methods with an expanded reference set. *Anal Biochem* 287:252–260
11. Fiedler S, Cole L, Keller S (2013) Automated Circular Dichroism spectroscopy for medium throughput analysis of protein conformation. *Anal Chem* 85:1868–1872
12. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637
13. Provencher SW, Glockner J (1981) Estimation of globular protein secondary structure from circular dichroism. *Biochemistry* 20:33–37
14. Van Stokkum LHM, Spoelder HJW, Bloemendal M et al (1990) Estimation of protein secondary structure and error analysis from CD spectra. *Anal Biochem* 191:110–118
15. Siligardi G, Panaretou B, Meyer P et al (2002) Regulation of Hsp90 ATPase activity by the co-chaperone Cdc37p/p50cdc37. *J Biol Chem* 277:20151–20159
16. Siligardi G, Hussain R (2010) Lindon J, Tranter G, Koppelaar D (eds) In: *Applications of circular dichroism encyclopedia of spectroscopy and spectrometry*, 2nd ed. vol 1. Elsevier, Oxford. pp 9–14
17. Martin SR, Schilstra MJ, Siligardi G (2011) Chapter 7: circular dichroism. In: Podjarny A, Dejaegere A, Kieffer B (eds) *Biophysical approaches determining ligand binding to biomolecular targets, detection, measurement and modelling*. RSC Publishing, Cambridge, pp 226–246
18. Prodromou C, Siligardi G, O'Brien R et al (1999) Regulation of Hsp90 ATPase activity by tetratricopeptide repeat (TPR)-domain co-chaperones. *EMBO J* 18:754–762
19. Strickland EH (1974) Aromatic contributions to circular dichroism spectra of proteins. *CRC Crit Rev Biochem* 2:113–175
20. Kohda D, Morton CJ, Parkar AA et al (1996) Solution structure of the link module: a hyaluronan-binding domain involved in extracellular matrix stability and cell migration. *Cell* 86:767–775
21. Blundell CD, Mahoney DJ, Andrew A et al (2003) The link module from ovulation- and inflammation-associated protein TSG-6 changes conformation on hyaluronan binding. *J Biol Chem* 278:49261–49270
22. Stewart LM, Bakker EP, Booth IR (1985) Energy coupling to K⁺ uptake via the Trk system in *Escherichia coli*: the role of ATP. *J Gen Microbiol* 131:77–85
23. Bertrand T, Kothe M, Liu J et al (2012) The crystal structures of TrkA and TrkB suggest key regions for achieving selective inhibition. *J Mol Biol* 423:439–453

High-Throughput Studies of Protein Shapes and Interactions by Synchrotron Small-Angle X-Ray Scattering

Cy M. Jeffries and Dmitri I. Svergun

Abstract

Solution-based small angle X-ray scattering (SAXS) affords the opportunity to extract accurate structural parameters and global shape information from diverse biological macromolecular systems. SAXS is an ideal complementary technique to other structural and biophysical methods but it can also be applied alone to access structural information that is otherwise unobtainable using high-resolution methods. Macromolecular structures ranging from kilodaltons to gigadaltons can be analyzed, which encompasses the size of most proteins and functional cellular complexes. The SAXS analysis is performed using only a few microliters of solution containing microgram quantities of purified material in sample environments that can be tailored to mimic physiological conditions or altered to suit a particular question. High-brilliance synchrotron X-ray sources and parallel advances in hardware and computing have reduced data acquisition times to the millisecond range and the application of automated methods have allowed data processing and low resolution shape modelling to be completed within minutes. These developments have paved the way for high-throughput studies that generate significant quantities of structural information over a short period of time. Here, we briefly consider the basics of SAXS and describe major methods and protocols employed in high-throughput SAXS studies.

Key words Small-angle X-ray scattering, Solution scattering, Monodisperse systems, High-throughput data processing, Macromolecular shape determination, Automated methods, Structural genomics, Proteomics, Systems biology

1 Introduction

Of the battery of techniques available to structural biologists it is perhaps small-angle X-ray scattering (SAXS) that offers investigators the most conceptually straightforward and practical avenue to investigate the shapes of macromolecules in solution [1–5]. The principle is simple: (1) obtain X-ray scattering data from a dilute sample of non-interacting monodisperse proteins in solution and; (2) subtract scattering contributions made by the solvent in which

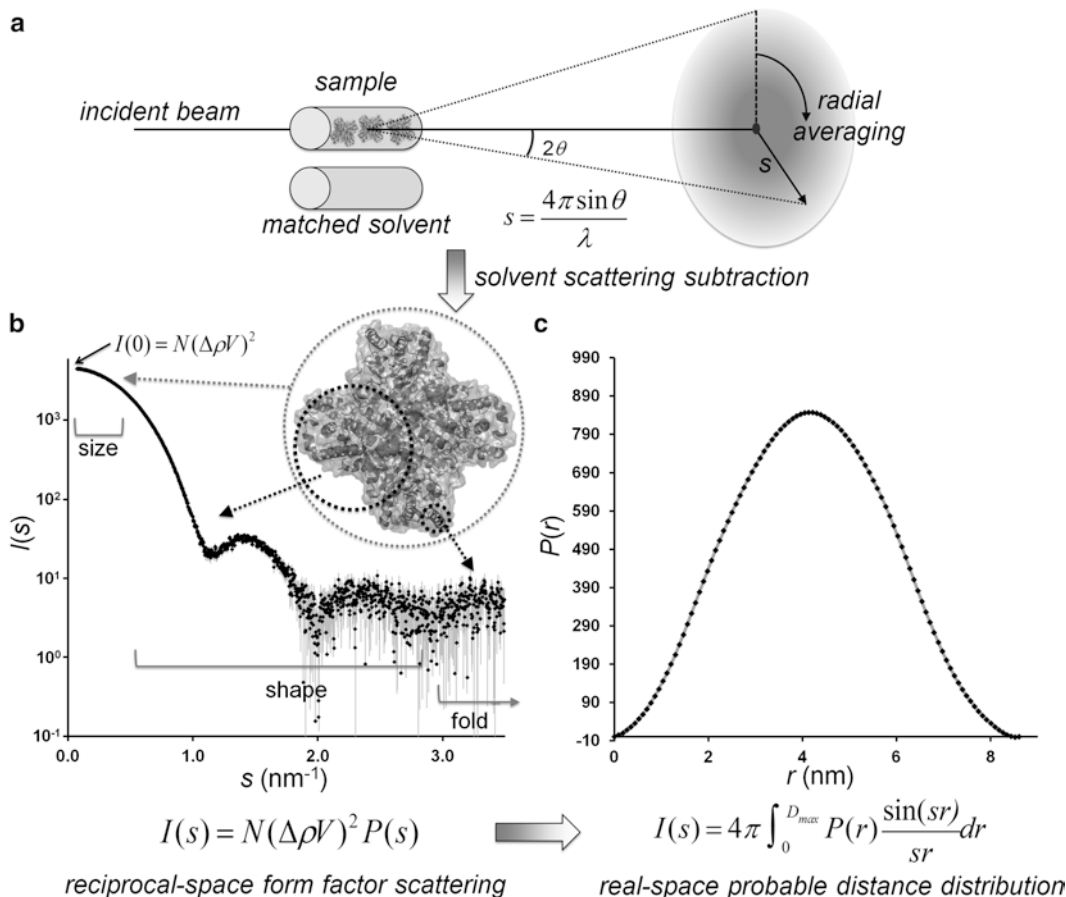


Fig. 1 The basics of small-angle X-ray scattering. **(a)** Schematic representing a basic SAXS experiment. A primary incident beam of collimated X-rays with a known λ illuminates a sample. The majority of X-rays are transmitted; however, a small proportion undergo elastic scattering events with the electrons of the sample and coherently interfere to produce scattering intensities as a function of angle (θ), that are recorded in terms of momentum transfer, s . Radial averaging of $I(s)$ vs s produces a 1D scattering profile. The subtraction of scattering contributions made by a matched background solvent produces a reduced scattering profile representing the scattering from macromolecules in the sample. **(b)** The reduced 1D scattering profile of monodisperse glucose isomerase. Highlighted are the main three regions of $I(s)$ vs s from which the shape of a single protein can be restored from the data, including intensities derived from all scattering centers bounded by the volume of the protein that are assessed via extrapolation to zero angle, $I(0)$. The size, shape, and fold of the protein are encoded at ever-increasing angle by the form factor, $P(s)$; however, the magnitude of $I(s)$ at high angle is orders of magnitude less intense, making it exceptionally difficult to accurately assess “high-resolution” details. **(c)** Real-space distance distribution of glucose isomerase, $P(r)$ vs r . The probable distribution of atom-pairs within the protein volume in real-space relate to $I(s)$ vs s measured in reciprocal space via an inverse Fourier transform

the protein is suspended. The resulting background-subtracted scattering intensities measured as a function of angle (usually small angles less than 3°) reflect the distribution of correlated distances between scattering centers within a protein from which the shape can be restored (Fig. 1).

It has now become routine to model the shapes of macromolecules, their complexes or assemblies against SAXS data [6–8] and, increasingly, SAXS is being used to probe macromolecular ensemble states and the formation of complexes or assemblies in real time [9–11]. Indeed, the ever-increasing number of scientific publications in the literature [12] suggest that SAXS is an invaluable tool for structural biologists [13] and is opening up new avenues of experimental inquiry that have been otherwise closed to NMR spectroscopy or X-ray crystallography. As a point of future interest, a bioinformatic analysis of the human genome indicates that of the 20–25,000 or so proteins are encoded by our DNA [14], 40 % contain significant regions of structural disorder and 25 % are completely intrinsically disordered [15, 16]. Aside from challenging the traditional paradigm of how protein structure relates to function [17], this statistic poses a considerable challenge for structural biologists. Disordered or partially disordered proteins do not readily crystallize for X-ray diffraction structural investigations and NMR spectroscopy structure determination becomes inordinately more difficult, especially as protein targets increase in size beyond 30 kDa. The appeal of SAXS is that it can be applied to the analysis of almost any macromolecular system—disordered or not—to generate meaningful structural information that can be used to develop models representing the global solution state, or states, of proteins, including how these state(s) respond to changes in solution conditions [18–25].

2 The Basics of SAXS

The principle of small-angle scattering has been well described over the course of nearly a century of inquiry into the nature of light and matter and has been used as an established technique across scientific disciplines that span chemistry, biology, physics, geology, material science, and engineering. Advanced and detailed explanations can be found in monographs [26] and [27, 28] including a recent book devoted to biological macromolecules [29] as well as in comprehensive reviews [30–32]. Here, we briefly present the basics of the technique; those interested in more detailed treatment are referred to the above literature.

SAXS is based on the elastic scattering of X-ray photons by electrons. Any electron in the path of an X-ray beam has the potential to scatter X-rays—electrons from buffer components, water, sample holder (capillary), and of course protein. Proteins comprising a sample have relatively time-preserved and correlated atomic distances bound within the volume (V) of a molecular envelope. Consequently, scattered X-rays from within each individual protein will constructively interfere and sum to produce scattering at low angles. For the entire sample, often comprising a population of

isotropically tumbling non-interacting identical particles, i.e., a monodisperse solution, the intensity gets time- and rotationally averaged (Fig. 1a). The magnitude of the total signal depends on the number-density of particles (N , concentration) while the way in which the intensity decreases as the scattering angle increases—or form factor—depends, in part, on the frequency of distances between the electron scattering centers within the volume boundary, i.e., the shape. Longer distance correlations and overall molecular size are represented at the lowest angles, while as the angle increases ever-shorter distances are represented in the scattering profile. However, the time- and rotationally averaged nature of the scattering data limits the information content: very short distances representing, for example, the fold of a protein scatter very weakly at high angle, making it exceptionally difficult to extract meaningful “high resolution details” (Fig. 1b). When applied to structural biology, SAXS directly provides global structural parameters including molecular weight (MW), radius of gyration (R_g), maximum dimension (D_{\max}) as well as volume and the probable atom-pair distance (r) distribution ($P(r)$ vs r) in real space that can be used to restore the overall shapes of proteins in solution (Fig. 1c).

Small molecules (e.g., salts) dissolved in water, and water itself, that otherwise lack time and long-range distance correlations produce flat scattering in the small-angle regime. However, small molecules, water (and the sample capillary), all contribute to the total scattering and their contributions have to be subtracted to “reveal” the scattering from the protein alone, i.e., SAXS is a subtractive technique:

$$\text{SAXS}_{(\text{target})} = \text{SAXS}_{(\text{target}+\text{solvent}+\text{cell})} - \text{SAXS}_{(\text{solvent}+\text{cell})}. \quad (1)$$

It is important when collecting SAXS data to ensure that samples and backgrounds are collected under identical conditions (e.g., temperature, pressure, exposure time) and that the background solvent used for subtraction has a matched chemical composition as the protein sample. Furthermore, the average number of electrons per unit volume of a protein, or scattering length density ρ , has to be different to that of the supporting solvent, ρ_s (i.e., the contrast $\Delta\rho = \rho - \rho_s \neq 0$), otherwise the “scattering boundary” between the protein and solvent will be effectively rendered invisible, i.e., the background will scatter as intensely as the sample and any SAXS profile effectively nullified. Proteins contain predominantly light atoms, and the average protein electron density is about 420 e/nm³, whereas that of water is 334 e/nm³, such that the difference signal from a dissolved protein in solution is quite low. Consequently, and because the contrast is further diminished when adding electron-rich compounds like salts or glycerol to the background solvent, it is often the case that protein-SAXS is performed in aqueous buffers where excessive quantities of additives are avoided so as to maintain the contrast of the system.

A monochromatic beam of X-rays of a known wavelength λ (typically, about 0.1 nm) can pass directly through a sample, absorb or scatter. The transmitted beam measurement allows one to account for the absorption by the sample, while the scattered X-ray intensities (I) are collected by an X-ray detector (e.g., a single-photon-counting hybrid pixel detector [33]). To obtain improved data statistics, two dimensional (2D) detectors are employed. As most protein samples under dilute conditions tumble randomly in solution, the resulting 2D-scattering pattern will be isotropic and can be radially averaged across the detector. This radial averaging produces a 1D-SAXS profile of scattering intensity expressed as a function of angle, specifically the momentum transfer, s :

$$I(s) \text{ vs } s, \quad s = \frac{4\pi \sin \theta}{\lambda}, \quad (2)$$

where θ is half the angle between the incident beam and the scattered radiation (Fig. 1a). For monodisperse systems consisting of non-interacting randomly oriented identical particles, the background-corrected intensity $I(s)$ is proportional to the scattering from all particles, averaged over all orientations:

$$I(s) = N \left\langle \left(\int_V (\rho(\vec{r}) - \bar{\rho}_s) e^{i\vec{s} \cdot \vec{r}} d\vec{r} \right)^2 \right\rangle_\Omega \quad (3)$$

where N is the number of dissolved scattering particles in the irradiated volume and $\langle \rangle$ indicates that the net scattered intensity emanating from all scattering centers within each particle, i , is rotationally averaged over all orientations (Ω). The integration is performed within the particle volume V , and each atom pair inside the particle gives rise to a scattered circular wave whose form is expressed as $e^{i\vec{s} \cdot \vec{r}}$, where \vec{r} is the vector between atom pairs. The amplitude of the scattering from each atom pair is proportional to the product of the contrast values at each atom ($\rho - \rho_s$) and the total scattering is the sum of all such contributions and here is expressed as an integral over the total particle volume.

2.1 $I(s)$ as It Relates to Particle Volume, Contrast, Form Factor, and Molecular Weight

When taking the factors described in Eq. 3 into account, $I(s)$ can be expressed as:

$$I(s) = \sum_i^n \left[(\Delta\rho_i V_i)^2 P_i(s) \right] S(s), \quad (4)$$

i.e., the summed contribution of the scattering from each individual particle within a sample weighted by the contrast against the solvent and volume squared multiplied by the form factor, $P_i(s)$ that represents “within-particle” scattering in reciprocal space that is dependent on the shape of the protein in real space. The structure factor term, $S(s)$, describes the influence on the scattering

caused by correlated distances of closest approach between each individual within the population, or “between particle scattering.” If a sample is sufficiently dilute and monodisperse, i.e., consists of N identical non-interacting particles, $S(s)$ will limit to unity and the above relationship simplifies to:

$$I(s) = N(\Delta\rho V)^2 P(s) \quad (5)$$

and thus $I(s)$ will be proportional to the form factor, $P(s)$, of a *single* protein in solution. At zero angle ($s=0$) the magnitude of $I(s)$ will primarily depend on the number of scattering centers within the bound squared-volume of the protein—independent of the shape—weighted by the protein concentration and contrast squared, i.e.,

$$I(0) \approx N(\Delta\rho V)^2. \quad (6)$$

Consequently, if the magnitude of the scattering at zero angle can be assessed and the data placed on an absolute scale (in units of cm^{-1}), e.g., by measuring water scattering [34] then the molecular weight (MW) of a protein can be determined:

$$\text{MW} = \frac{I(0) N_A}{c(\Delta\rho v)^2}, \quad (7)$$

where c is the concentration (g/cm^3), v is the partial specific volume (cm^3/g) of the protein and N_A Avogadro’s number. As there is no way to directly measure $I(0)$ because the primary beam used to illuminate the sample is coincident with $s=0 \text{ \AA}^{-1}$, $I(0)$ is usually determined via extrapolation using the Guinier approximation [35] or from the real-space atom-pair distance distribution (see below). Of particular importance, as $I(0)$ can be used to extract MW information from a scattering profile, its determination from a SAXS experiment is important to assess the sample monodispersity, which is key to further SAXS data modelling and interpretation [2].

The MW of a scattering species can also be estimated by comparing $I(0)$ with a secondary standard of monodisperse particles, for example scattering by a known protein (e.g., lysozyme, bovine serum albumin, cytochrome-C [36]) with known concentration (in mg/mL). From the relation,

$$\text{MW}_{\text{standard}} \propto \frac{I(0)_{\text{standard}}}{c_{\text{standard}}} \quad (8)$$

a constant, K , can be determined:

$$K = \frac{I(0)_{\text{standard}}}{c_{\text{standard}} \cdot \text{MW}_{\text{standard}}}. \quad (9)$$

and the MW of the protein can be calculated as:

$$\text{MW}_{\text{protein}} = \frac{I(0)_{\text{protein}} \cdot c_{\text{standard}} \cdot \text{MW}_{\text{standard}}}{c_{\text{protein}} \cdot I(0)_{\text{standard}}}. \quad (10)$$

The advantage of this method is that there is no need to place the scattering data on an absolute scale to obtain $I(0)$, but there is an assumption that a target has a similar scattering length density as the secondary standard. For targets having a different scattering length density or perhaps more than one region of scattering density, e.g., a RNA/protein complex, a correction factor due to the different contrasts has to be determined.

2.2 Shape Information

The scattering intensity $I(s)$ —and thus the associated form factor in reciprocal space—relates to an atom-pair distance distribution function of the particle $P(r)$ in real space by a Fourier transform:

$$I(s) = 4\pi \int_0^{D_{\max}} P(r) \frac{\sin(sr)}{sr} dr \quad (11)$$

$$P(r) = \frac{r^2}{2\pi} \int_0^{\infty} s^2 I(s) \frac{\sin(sr)}{sr} ds, \quad (12)$$

where D_{\max} is the maximum dimension of the particle [37–41]. The summed-total of each distance frequency across the distribution, or the area under $P(r)$ vs r , relates to the contrast-weighted total number of scattering centers, or $I(0)$:

$$I(0) = 4\pi \int_0^{D_{\max}} P(r) dr. \quad (13)$$

Overall, the basics of macromolecular SAXS are embedded within a deceptively simple concept: a protein will have x -number of scattering centers distributed within its volume that are distance-correlated over time. The isotropic scattering intensities produced as a function of s from a monodisperse population of randomly tumbling proteins relate to the distribution of these distances and it is from this that structural parameters and shape information in real space can be extracted from the data that represent the size and shape of a single protein in solution.

3 Structural Parameters

In addition to the MW (calculated from $I(0)$), the R_g and volume of a protein can be obtained from relatively simple transformations of the data. The D_{\max} , or the maximum extent of the longest vector

length between atom-pairs, can be estimated from $P(r)$ vs r . Indeed, as Eq. 11 indicates, D_{\max} is a parameter that is selected to solve the Fourier transform fit to the data that produces the best $P(r)$ vs r model.

3.1 The Radius of Gyration

The R_g value equates to the contrast weighted root mean square (or quadratic mean) distance of all volume elements occupied by scattering centers with respect to the center of the scattering length density of a particle. The R_g is thus influenced by both the size and shape of the protein under investigation and generally provides information on the shape-weighted volume distributed around a proteins center of mass.

Guinier [35] showed that the scattering intensity at low angles is dependent on the radius of gyration as:

$$I(s) = I(0) e^{\frac{-s^2 R_g^2}{3}}. \quad (14)$$

The Guinier approximation holds true for most monodisperse globular protein systems in the interval defined by the condition $s_{\min} R_g < 1.3$ and offers an exceptionally useful transformation of the data to obtain $I(0)$ and R_g as well as providing an almost-immediate tool to assess sample quality. For monodisperse proteins, a plot of the natural logarithm of the experimental $I(s)$ vs s^2 —or Guinier plot—will produce a negative linear relationship in the range $s < s_{\min}$. Extrapolation to the $\ln I(s)$ intercept at $s=0$ will yield $I(0)$ and the slope of the plot will be proportional to R_g^2 . Observable upward deviations in $\ln I(s)$ away from linearity within the Guinier regime is an indication that the sample is either contaminated with larger species or contains nonspecific aggregates (i.e., $S(s)$ is positive). Repulsive interparticle interference caused by coulombic repulsion of surface charges between proteins, produces a decrease in intensity at low- s values resulting in a downward curvature away from linearity in Guinier plots (i.e., $S(s)$ is negative).

The R_g can also be determined from $P(r)$ vs r :

$$R_g^2 = \frac{\int_0^{D_{\max}} P(r) r^2 dr}{2 \int_0^{D_{\max}} P(r) dr} \quad (15)$$

i.e., R_g^2 corresponds to the second moment of $P(r)$. The determination of R_g , or $I(0)$ from $P(r)$ vs r (Eq. 13) can afford a more accurate assessment of these parameters compared to the Guinier approximation as the determination these structural parameters is based on fitting most of, or indeed all of the SAXS data to high- s . A Guinier plot on the other hand uses only the very lowest s -values from an entire dataset on which to base the extrapolation of R_g or $I(0)$. Consequently, the number of data points

used for the Guinier extrapolation is heavily influenced by particle size and shape. For example, lysozyme, a small 14 kDa globular protein, has a very reasonable Guinier region that extends to $s \approx 0.9 \text{ nm}^{-1}$ that on modern detectors measuring between $0.04 < s < 4 \text{ nm}^{-1}$ encompasses the first $\sim 20\%$ of data points. The Guinier region of BSA, a 66 kDa globular protein, extends to $s \approx 0.4 \text{ nm}^{-1}$, which is again reasonable for the Guinier extrapolation, using the first 10 % of the data. However, the Guinier region of GroEL, a large 800 kDa protein complex is limited only to $s \approx 0.2 \text{ nm}^{-1}$. Even on modern instruments, large proteins like the GroEL complex might have Guinier regions that typically encompass the first 2 % of points of the whole dataset. Complicating Guinier analysis is that as the shape of proteins move away from globularity, the Guinier approximation as linear out to $sR_g \sim 1.3$ begins to breakdown. Highly extended or filamentous proteins can have standard Guinier regions for values of $sR_g < 0.8\text{--}1.0$ depending upon the degree of asymmetry and this further reduces the number of low- s data points on which to base the extrapolation. For example, the modular, N-terminal fragment of cardiac myosin binding protein C, even though it has a similar MW as BSA of 70 kDa, has a low- s Guinier region that is limited to $sR_g < 1.0$ due to its structural anisotropy resulting in a very limited Guinier region encompassing the first 2 % of the total number of data points (Fig. 2, [42]).

Obtaining R_g and $I(0)$ from $P(r)$ vs r overcomes the problem of limited data points in the Guinier region for large or extended protein samples. However, in most instances both the Guinier approximation and $P(r)$ vs r estimates of R_g and $I(0)$ are very robust and will produce almost identical values for both structural parameters when used in combination for the analysis of SAXS data.

3.2 Porod Volume and Kratky Plot

The determination of MW from $I(0)$ requires an accurate assessment of the concentration of a protein in solution (Eq. 7) that in itself can be difficult to determine especially when the UV or visible absorption extinction coefficients (ϵ) used to calculate protein concentration are unknown or negligible (via the Beer–Lambert relationship, $\text{Abs} = \epsilon lc$). An alternative concentration-independent estimate of MW is based on the volume of a protein in solution. Porod [43] showed that for uniform particles with sharp boundaries the excluded volume V_p can be calculated as:

$$V_p = \frac{2\pi^2 I(0)}{Q}, \quad (16)$$

where Q is a Porod invariant or the area under a plot of $I(s)s^2$ vs s calculated to $s = \infty$,

$$Q = \int_0^\infty s^2 I(s) ds. \quad (17)$$

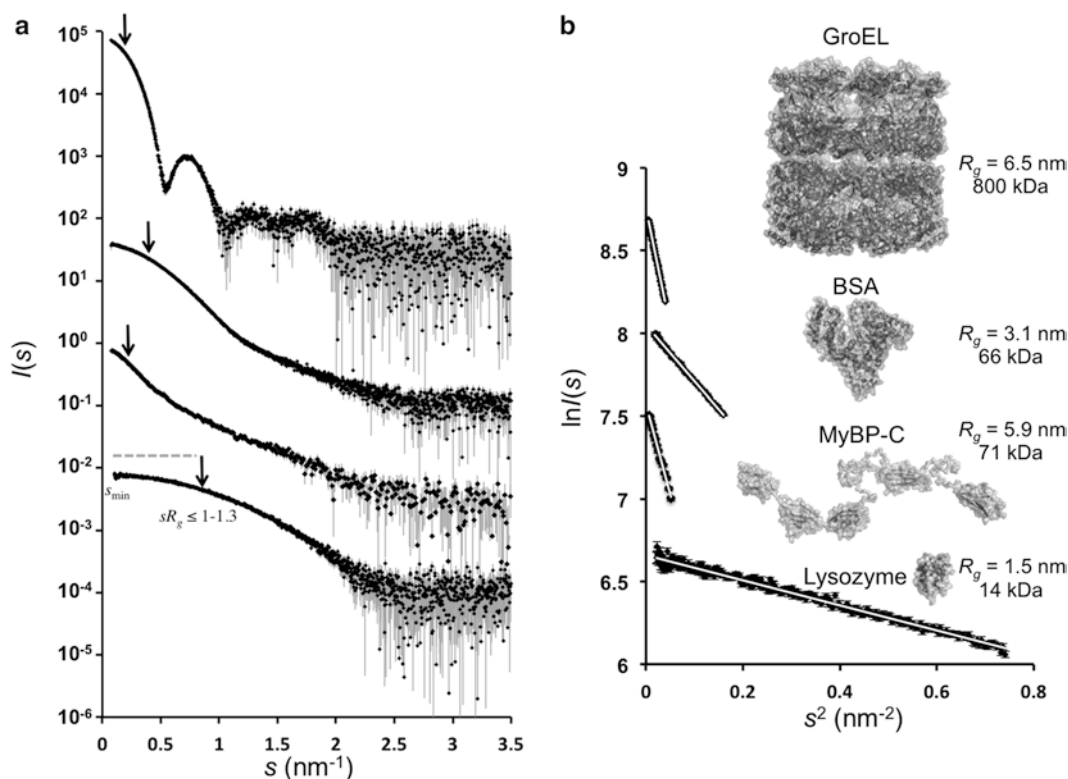


Fig. 2 The effects of size and shape on SAXS scattering profiles. (a) Scattering data from GroEL, bovine serum albumin (BSA), the N-terminal domains of myosin binding protein-C (COC4) [42] and lysozyme are represented, from *top* to *bottom*, respectively. In general, $I(s)$ vs s decays more rapidly for larger macromolecules. (e.g., compare the 800 kDa GroEL complex to the 14 kDa lysozyme.) The extent of the Guinier region for each protein is shown by *arrows*. (b) Guinier plots, $\ln I(s)$ vs s , of the same proteins showing that both the size and shape affects the linear extent of the relationship at very-low angle. It is from the slope of the plots that R_g can be determined. Note, for presentation purposes, $I(s)$ vs s have been scaled on the y -axis

Although proteins do not have uniform internal electron density, and although it is not possible to measure $I(s)$ to infinite s or collect SAXS data as a continuous function of s , (but rather as discrete Δs intervals) it is still possible to employ the Porod relationship in many practical applications. An apparent Q can be calculated from $I(s)s^2$ vs s , or Kratky plot, via:

$$Q' = \sum_{s_{\min}}^{s_{\max}} s^2 I(s) \Delta s \quad (18)$$

The application of empirical correction factors [44] or the subtraction of an appropriate constant from the scattering data [45] can provide a reasonable approximation of Q from Q' that relates the scattering of a protein to the scattering of a corresponding homogeneous body from which V_p can be determined. The V_p of a protein in nm³ is typically 1.5–2 times the MW in kilodaltons (kDa).

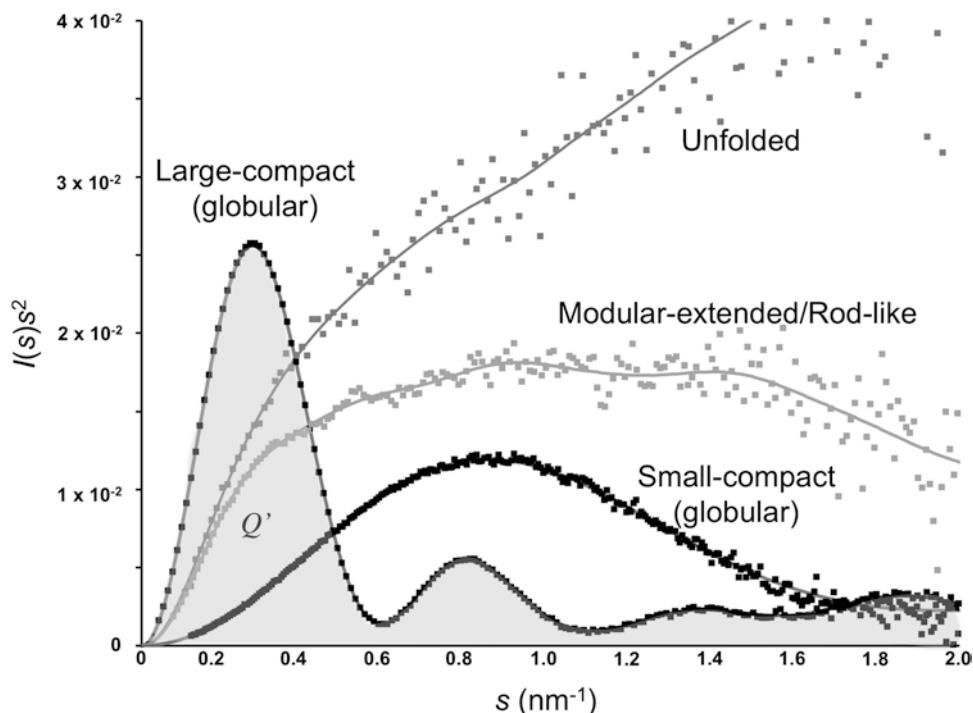


Fig. 3 Kratky Plots. Scattering data when plotted as $I(s)s^2$ vs s , or a Kratky plot, can be useful to obtain a value of Q' , or Porod invariant, from which the Porod volume and subsequent molecular weight of a protein can be evaluated. Furthermore, a qualitative assessment of the plot can reveal something about the shape, compactness or flexibility of a protein. Compact proteins produce a characteristic maximum, the position of which relates to R_g , while completely flexible/disordered proteins display a continual increase in $I(s)s^2$ as s increases. Severe deviations from globularity can affect Q' and the calculation of the Porod volume. Shown here are scaled Kratky plots from SAXS data collected from GroEL (large globular), lysozyme (small globular), the N-terminal domains of myosin binding protein-C (modular-extended/rod-like) [42], and completely denatured GroEL (unfolded)

However, caution—as always—must be applied when dealing with highly anisotropic or highly flexible/disordered proteins or systems that are “atypical” due to the presence of binding partners with different scattering length-densities (e.g., a protein–DNA complex) that have the effect of altering the empirical correction factors. In the case of flexible, or rod-like proteins, the decay in scattering intensities at high angle deviates sufficiently from Porod’s law that the estimation of Q' as it relates to Q will incur errors in the volume estimation (Fig. 3). However, when used astutely—and in parallel with $I(0)$ -based analysis, that is shape independent—the use of V_p provides a robust complementary way to assess the MW of globular particles [46].

3.3 Calculation of Distance Distributions

Due to the discrete properties of experimental SAXS data and the limited experimental s -range through which it is measured, the direct determination of $P(r)$ vs r from $I(s)$ using Eq. 12 is not possible. Therefore, indirect Fourier transformation methods [40, 41]

are usually employed to generate a $P(r)$ function that fits the experimental data. Here, $P(r)$ is described in terms of a linear combination of coefficient-weighted K orthogonal functions, $\varphi(r)$ (for example cubic B-splines) across the interval of $0-D_{\max}$ whose individual Fourier transforms $\psi(s)$ are known. The experimental data is presented as a coefficient-weighted sum:

$$P(r) = \sum_{k=1}^K c_k \phi_k(s) \quad (19)$$

The fit to the scattering profile is optimized by the coefficients, c_k , that multiply each $\varphi(r)$ so as to minimize the discrepancy (χ^2) between the experimental and calculated data while maintaining a smooth $P(r)$:

$$\Psi = \chi^2 + \alpha P(p) \quad (20)$$

The discrepancy, or goodness of fit,

$$\chi^2 = \frac{1}{N-1} \sum_{j=1}^N \left[\frac{I_{\text{exp}}(s_j) - c I_{\text{calc}}(s_j)}{\sigma(s_j)} \right]^2, \quad (21)$$

is a measure within the experimental error, σ , of the difference between every j th point of the experimental data set (I_{exp}) and the modelled intensity at every j th point. The penalty term, $P(p)$ is effectively a “smoothing term” for $P(r)$:

$$\alpha P(p) = \int_0^{D_{\max}} \left[\frac{dp}{dr} \right]^2 dr. \quad (22)$$

This term is required because solving Eq. 11 is a so-called ill-posed problem and would have provided unstable $P(r)$ functions if only goodness of fit is minimized. The regularization parameter $\alpha > 0$ acts as a balance between the stability of the computed $P(r)$ in real-space and the goodness of fit in reciprocal-space. If α is too small, the $P(r)$ fit to the data fit will be overly influenced by the discrete Δs point-to-point variation and the experimental error at each point and consequently the $P(r)$ vs r will be affected by too many oscillations. Conversely, if α is too high, the smoothness of the transform is prioritized over the fit to the data and the discrepancy, χ , will be compromised. In order to simplify the search for an appropriate solution, the program GNOM [37–39] automates the search for the number of coefficients and the α -weighting so that a solution is found where χ is minimized and either a single or series of stable $P(r)$ solutions are identified where the oscillations in the coefficients are diminished. The resulting output file from the program contains the best-fit real-space atom pair distance distribution

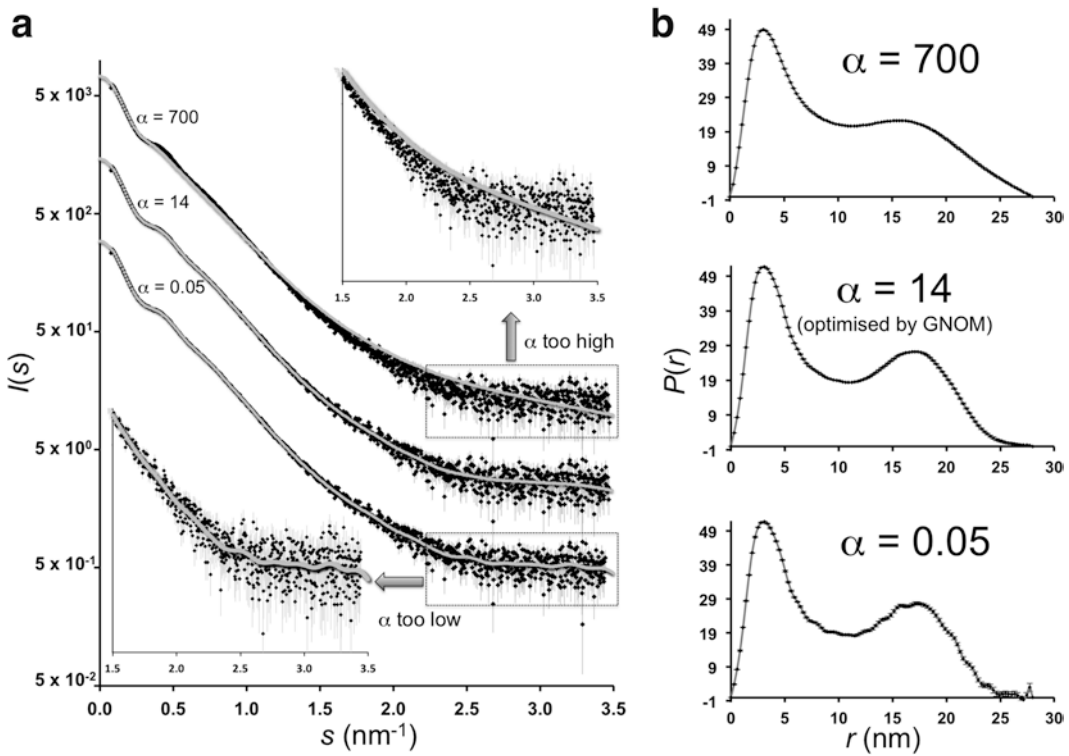


Fig. 4 Modelling $P(r)$ vs r fits to SAXS data. (a) The effect of altering the regularization parameter, α , on the $P(r)$ vs r fit to the scattering data and the stability of the resulting $P(r)$ vs r profiles (b) determined for maltose-binding protein fused to an extended coil-coil protein. If α is too large, the calculated $P(r)$ vs r , although smooth/stable, does not fit the data, whereas if α is too low, over-fitting of the data occurs generating to $P(r)$ vs r instability. The program *GNOM* uses a set of perceptual criteria to select an optimized value of α [37–39]

using the D_{\max} chosen to solve the indirect Fourier transform, as well as the smoothed $P(r)$ fit to the data that can be used as input for ab initio shape restoration and model building (Fig. 4).

4 Ab Initio Shape Restoration Using a Dummy-Atom Model (DAM) Approach

Modelling a three-dimensional (3D) shape of a protein derived from a 1D scattering pattern representing a rotationally and time-averaged sample is not trivial due to the loss of spatial resolution caused by spherical averaging. In 1970, Stuhmann showed that the information content of a SAXS profile can be conveniently described in terms of a sum of spherical harmonic functions [47, 48]. The application of spherical harmonics to model SAXS data has led to subsequent development of rapid analytical techniques to calculate SAXS curves from known structures, for example geometric shapes, volumes packed with dummy atoms, and high-resolution atomic structures, that have also been applied to modelling the shapes of proteins in solution [49–54].

Several approaches have been proposed for ab initio shape restoration [49, 53, 55–60], which is now a routine way of modeling the shape of monodisperse protein samples against solution-SAXS data. The algorithms implemented in the program *DAMMIN* [49] generate possible 3D dummy atom models (DAMs) of a scattering particle (assumed to be of a uniform scattering density) from a 1D scattering profile. Using a simulated annealing (SA) protocol and the smoothed $I(s)$ vs s profile derived from the $P(r)$ model fit to the experimental data as a target function, *DAMMIN* starts by packing a sphere of a defined maximum dimension, usually D_{\max} , with an ensemble of smaller spheres (dummy atoms) that are assigned a “phase”: e.g., solvent (0) or protein (1). The configuration of the DAM ensemble is represented as a binary string of length M (the total number of small spheres within the DAM), or “phase assignment vector.” A model is characterized by an energy function, which, similar to Eq. 20, contains a discrepancy term and a penalty for non-compactness and disconnectivity of the model. At the beginning of the SA, when the temperature of the system is high, there is only a small energy difference between the phase assignments for any one dummy atom within the ensemble. As the system is cooled, each individual bead of the model is changed; the theoretical scattering pattern is computed using spherical harmonics, and the discrepancy between the scattering from the bead model and the experimental SAXS data is calculated. Eventually the energy difference between the two phases at any one dummy atom position within in the entire DAM ensemble increases to such an extent that the phase at a position becomes “fixed” as either solvent or protein. The net result, after the assigned solvent phase is removed, is a compact interconnected DAM that represents the shape of a protein in solution that fits the scattering data (Fig. 5).

A faster version of *DAMMIN* called *DAMMIF* [53] greatly reduces the computational time necessary to restore a shape of a protein against a scattering pattern. Unlike *DAMMIN* that starts with a fully randomized closed search volume of dummy atoms, *DAMMIF* begins with filling an isometric volume with the same R_g determined from an experiment and implements an unconstrained volume search that can grow during the SA procedure. By rejecting disconnected solutions and only taking into account those dummy atoms that contribute to the scattering for the first time at each annealing step (as opposed to all solutions and all atoms), and by keeping and reusing all the computed contributions from the dummy atoms, *DAMMIF* speeds up the calculation of scattering amplitudes of the DAM by a factor of 20–25 compared to *DAMMIN*. A fast *DAMMIN* run takes less than a minute on a desktop PC.

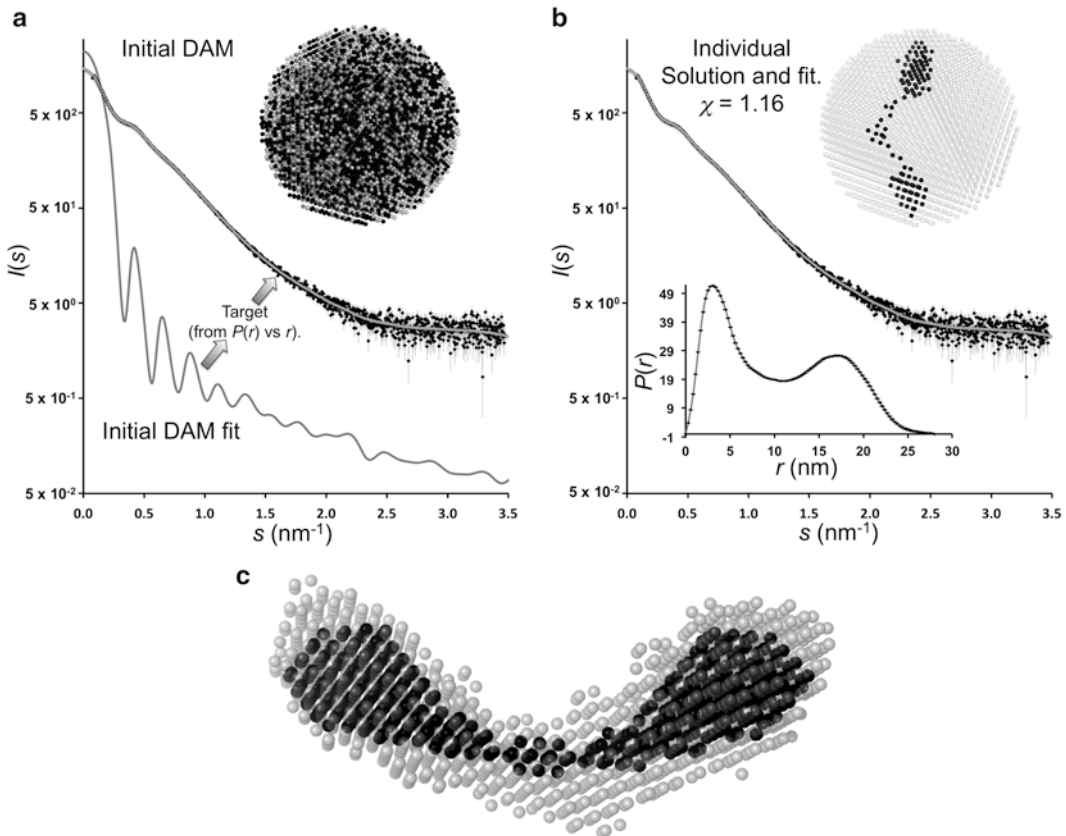


Fig. 5 Ab initio dummy atom modelling (DAM). **(a)** The program *DAMMIN* uses $P(r)$ vs r outputs (e.g., from *GNOM*) as an input for dummy atom modelling. The initial DAM consists of multiple beads randomly assigned as two phases: “protein,” black spheres and solvent, grey spheres. **(b)** After slow-cool annealing, a model is generated (*black spheres*) that fits the SAXS data representing an individual DAM solution. **(c)** *DAMMIN* is run multiple times to generate several individual models that are subsequently aligned to evaluate the consistency of the individual solutions (*grey spheres*) and then assessed in terms of spatial occupancy and volume to generate an average representation of the scattering particle (*black spheres*). In this instance, the ab initio shape reconstruction of maltose-binding protein fused to an extended coil-coil protein is shown to illustrate the process of obtaining the shape of a protein from SAXS data

4.1 Model Convergence

DAMMIF offers a rapid and robust modelling program that is well suited for the initial evaluation of the shapes of proteins in solution especially at high-throughput synchrotron SAXS facilities where results can be computed in almost real time. However, more than one 3D shape can theoretically fit a 1D scattering profile and therefore it is necessary to evaluate multiple ab initio solutions to assess convergence towards an average solution [61]. *DAMMIF* or *DAMMIN* are usually run multiple (at least 10) times after the initial shape restoration and the resulting solutions aligned, averaged and volume/occupancy corrected to produce an average model of the scattering particle [61, 62] (Fig. 5). The averaging procedure (implemented in the programs *DAMAVAR* [61]) produces, among

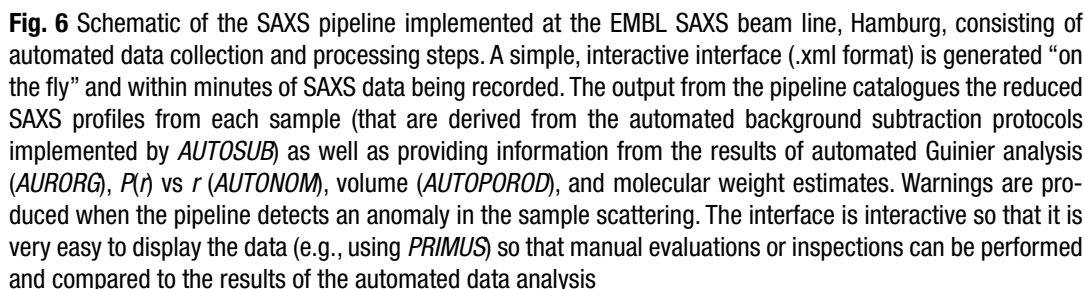
other parameters, a normalized spatial discrepancy (NSD [62]) value that is helpful in assessing how similar the individual restorations are and how unique the solution might be. In general, either the averaged or the most typical model is selected, i.e., a model with the lowest NSD relative to the other individual solutions. An alternative program, *DAMCLUST* [45] clusters the generated models into spatially distinct groups then compares the NSD between clusters to identify whether any viable alternative shapes are indeed represented in the SAXS profile (this program is especially useful to analyze symmetric models or structures constructed by rigid body modelling).

5 Automation

The approaches described in Subheadings 1–4 provide a background for the high-throughput analysis of SAXS data and the determination of macromolecular shapes. The basic steps for an automated SAXS “pipeline” [63] and an essential display output is summarized below and shown in Fig. 6:

- Measurement of SAXS data from a buffer or background solvent.
- Measurement of SAXS data from a protein sample.
- Radial averaging of isotropic 2D SAXS data to produce 1D scattering profiles.
- Subtraction of background 1D scattering profiles from the 1D scattering profile of the sample.
- Determination of Guinier region and $P(r)$ vs r .
- Assessment of molecular weight of a protein based on $I(0)$ or volume.
- Determination of R_g from Guinier or $P(r)$ vs r .
- Determination of an initial ab initio shape.

Each step in the process of structural characterization a protein in solution is amenable to high-throughput automation and such automation has been implemented at a number of synchrotron SAXS beam lines [63–69]. Automated sample loading [64, 70] and rapid (down to millisecond range) acquisition times—thanks to the high brilliance of synchrotron X-ray sources—allows hundreds of protein samples to be studied in diverse conditions in a very short period of time [71]. Yet, dealing with thousands of datasets can be overwhelming and can cause a major data processing bottleneck. The latter is relieved by automated methods that provide rapid access to the data and, if statistical tests have been incorporated into the automation process, the provision of immediate feedback on the state of a sample, for example alerts flagging



miss-matched buffers or sample aggregation. Thus, with automated data reduction, processing, and modelling in place [63, 68, 72], the emphasis of an experiment at a SAXS beam line shifts toward the production of high-quality samples or tailoring sample conditions to optimize the scattering results. Automation allows for on-the-spot decisions to be made and time to adjust the sample parameters of an experiment. Importantly, automated methods should be flexible and cater for a diverse spectrum of users including novices and non-specialists who want to use SAXS to complement already existing structural biology programs. Given the requests from the community, several synchrotron sites have developed tools for the automation of SAXS data acquisition, reduction and interpretation [65–68]. In the following, we shall describe

perhaps the most comprehensive automated pipeline developed at the EMBL Hamburg for SAXS beam lines at the DESY synchrotron [63, 64]. This pipeline has been exported to other sites (e.g., ESRF, Grenoble [69]).

5.1 Automated 2D Data Integration

Rapid data processing begins by standardizing the parameters necessary to perform accurate 2D to 1D data integration thus ensuring that each dataset is scaled to the same frame of reference [63, 73]. The initial steps include defining the beam center and angular range for a particular instrument configuration, e.g., by measuring the scattering from a standard sample (e.g., silver behenate powder) [74]. Furthermore, the 2D radial averaging usually incorporates additional corrections for detector masked regions and sensitivity as well as normalization against transmitted sample intensities and exposure periods. Once standardized against the s -axis and normalized, the same integration parameters are applied to all subsequent 2D images to produce 1D profiles of s , $I(s)$ and the associated error on $I(s)$ in a format compatible for automated processing (e.g., in a simple columnar ASCII format [75] with appropriate header and/or footer information). At most synchrotron SAXS facilities the automated integration process takes seconds (for example using the program *RADDAVER*) and is performed on-line with the measurements. Most conveniently, the calibrations made for the standardized setup are valid for multiple groups and the users are not required to conduct any additional measurements or actions; the system can also be easily applied to recalculate the 1D profiles from the stored 2D images in a batch mode if necessary [63].

5.2 Automated Data Subtraction

Of particular importance for the interpretation of SAXS data is the accurate subtraction of background scattering contributions from the supporting solvent. Complicating this automated process is that proteins, in general, scatter X-rays very weakly, approximately one order of magnitude higher than the solvent scattering at very-low angle (down to a fraction of a percent at higher angles). Furthermore, proteins are often susceptible to radiation damage that for SAXS largely leads to X-ray induced aggregation under the intense fluxes encountered at synchrotron sources. Any automated background subtraction algorithms must have mechanisms in place to deal with these issues in real-time. The program *AUTOSUB* [73] performs buffer subtraction while incorporating statistical checks to ensure that the effects of minor beam instability or radiation damage are identified and taken into account. A typical collection strategy for *AUTOSUB* processing might consist of collecting multiple (e.g., 20×50 ms) data frames from the buffer and sample, followed by an additional 20×50 ms buffer measurement. Using header information from the 2D to 1D integration procedure, *AUTOSUB* automatically recognizes which data files correspond

to the background or sample scattering and assesses the stability of $I(s)$ vs s across each individual frame by applying a standard statistical F -test. If the individual data frames for each buffer or sample set are stable through the entire exposure period, *AUTOSUB* will average the corresponding buffer and sample frames and automatically subtract the background scattering from the sample scattering to produce an averaged 1D SAXS profile of the protein in solution (scaled by the protein concentration as provided by the User).

AUTOSUB also considers three other possible subtractions if the averaging process does not yield data with statistically stable frame-to-frame variances:

- Sample—buffer1.
- Sample—buffer2.
- Sample— $\frac{1}{2}$ (buffer1 + buffer2).

For each combination, statistical criteria are calculated that evaluate the magnitude of $I(s)$ at high- s —where $\Delta I(s)$ is expected to be small, i.e., limit to 0—to identify buffer over-subtraction that is flagged by the appearance of negative intensities at high angle. Furthermore changes in the pre- and post-sample buffer scattering intensities are calculated, for example to assess whether fouling of the sample capillary has occurred during exposure of the protein to X-rays. Importantly, an additional criterion is incorporated that monitors frame-to-frame changes in the R_g of the sample. The program *AUTORG* [73] works in parallel with *AUTOSUB* to produce a quality estimate of the least-squares linear fit to the Guinier region of the SAXS profiles (*see* next section). Overall, the automated subtraction procedure determines an overall scoring matrix so that discrepancies in buffer scattering intensities, buffer over-subtraction or systematic deviations in R_g (e.g., caused by X-ray induced aggregation) are flagged. Only the sample and buffer frames that produce the best composite score are selected as the final reduced 1D scattering profile. Although *AUTOSUB* will perform the subtraction process even if sample quality or stability is severely compromised, the warnings of the inherent problems encountered during subtraction, for example “no R_g found” or “too many negative intensities” will be produced. Such warnings indicate that severe aggregation problems have occurred in the sample or that there is a significant buffer miss-match.

5.3 Automated Determination of R_g and $I(0)$

Although the Guinier law (Eq. 14) appears a very simple and straightforward approach, automated R_g determination is not trivial given the limited validity of this approximation (condition $sR_g < 1.3$). As evident from the discussion in Subheading 3.1 and from Fig. 2, the Guinier range depends both on MW and on the particle anisometry such that no universal recipe exists to select the fitting interval. Until recently, determination of R_g was a largely interactive procedure. The first published automated approach implemented in a program

AUTORG performs a series of computations to extract the best estimate of R_g and $I(0)$ from a SAXS profile [63, 73]. In selecting for the most suitable s -range for the Guinier approximation *AUTORG* first defines a feasible possible range of search and scans for any severe deviations from linearity in the Guinier plot at the very low angles (e.g., caused by parasitic scattering near the incident beam). The program then performs an exhaustive “ sR_g interval search” by making Guinier fits in overlapping $[s_{\min}, s_{\max}]$ ranges and statistically assesses these fits in terms of the number of points fitted and deviations from linearity. The systematic differences in R_g between different $[s_{\min}, s_{\max}]$ ranges are evaluated yielding a overall quality rating of the statistically optimal Guinier region. Ultimately *AUTORG* produces an estimate of R_g and $I(0)$ based on the best fit to the selected Guinier region as well writing appropriate warning flags if there is something wrong with the sample (e.g., “possibly aggregated” if no stable R_g can be determined). If the configuration of an automated pipeline incorporates $I(0)$ derived from a secondary standard of known concentration (see Subheading 2.1, Eq. 8) and the scattering intensities of each protein sample are scaled to this standard, the $I(0)$ determined from *AUTORG* can be used to automatically estimate the MW of the protein [63, 73].

5.4 *P(r)* Function, Volume, Shape

The calculation of the real-space atom-pair distance distribution and of the Porod volume (Subheadings 3.3 and 3.2, respectively) are also amenable to automation. The program *AUTOGNOM* [73] calculates multiple solutions of the regularized indirect Fourier transformation fit against the scattering data at multiple D_{\max} (Eq. 11) using the perceptual criteria employed in *GNOM* [37] to generate multiple $P(r)$ vs r profiles. The D_{\max} scanned by *AUTOGNOM* span a $2R_g$ – $4R_g$ length interval, where R_g is determined by *AUTORG*. The $P(r)$ functions calculated for each D_{\max} (Eqs. 19–22) are scored and the final $P(r)$ vs r has the highest quality fit against the SAXS data combined with the condition that $P(r)$ decays smoothly to 0 as D_{\max} is reached (i.e., $P(D_{\max})$ and its first derivative approach 0). If the final R_g determined from this analysis is different to that calculated by *AUTORG* (this may occur especially for particles with a limited Guinier region), automated warnings can be set up to inform of these differences and the necessity for further inspection of the data. The *AUTOGNOM* output file is directly passed onto *DAMMIF* for ab initio shape restoration. As indicated in Subheading 4.1, shape determination is a rapid procedure and at least a single *DAMMIF* run can be performed in real time to provide the user with the shape estimate of the particle in-line with experimental data acquisition, even in a high-throughput mode.

As previously mentioned (Eqs. 16 and 17) the Porod volume, V_p , can be estimated from a scaled ratio of the contrast-weighted summed total scattering from each scattering center within a

particle, $I(0)$, and the Porod invariant Q [43]. There are obvious difficulties, however, in determining Q directly from an s -limited, discrete scattering dataset and correction factors need to be applied to estimate V_p (Subheading 3.2). The program *AUTOPOROD* [45] uses the value of $I(0)$ produced by *AUTORG* and computes Q from the smoothed $I(s)$ vs s scattering profile from *AUTOGNOM* and calculates V_p via:

$$V_p \approx \frac{2\pi^2 I(0)}{h \int_0^{s_{\max}} s^2 [I(s) - K] ds}. \quad (23)$$

Here, s_{\max} is defined to contain the portion of the scattering data where $I(s)$ diminishes to approximately two orders of magnitude less than $I(0)$, K is a correction factor that enforces an s^{-4} decay in the scattering intensities in that range (to account for the inhomogeneous scattering length density internal to a protein), and h is an empirical correction accounting for the finite range of integration [76]. For most globular, or near-globular proteins, the scattering intensity decays between s^{-4} and s^{-3} , and thus *AUTOPOROD* generates reasonable volume estimates that can be converted into a *MW* (e.g., if s is measured in nm^{-1} , division of V_p by 1.7 will produce a *MW* estimate in kDa). However, for disordered ($s^{-1.7}$), rod-like (s^{-1}) or disc-shaped (s^{-2}) proteins, the volume and consequent *MW* estimate can incur errors. In these—and indeed all—instances, the *MW* from V_p should be cross-checked against that obtained from secondary-standard concentration dependent $I(0)$ analysis (Eq. 10) or the *MW* estimated from the volume of *DAMMIF* ab initio shape reconstructions (where $\text{MW} \sim \text{DAMMIF volume}/2$). A well set-up automated SAXS-data analysis pipeline will report all three *MW* values that, when combined, can be highly informative for assessing the monodispersity of a sample and consequently the validity of the corresponding model.

6 Concluding Remarks

As the synthesis of structural biology with proteomics and systems biology continues to evolve, access to fast and reliable structural methods will become increasingly important to investigate diverse macromolecular interactions and the structural dynamics that underpin cellular processes. The automation of SAXS data acquisition, reduction and analysis, and the subsequent delivery of results within minutes of data collection, affords investigators an opportunity to obtain interpretable data from multiple samples across multiple sample environments in a very short period of time. SAXS is a unique structural biological technique for an efficient sampling structural space and it is the combined brilliance of synchrotron radiation

sources and the parallel development of automated methods that have converted this potential into a reality. However, although SAXS yields accurate structural parameters and shape information, it remains a low-information technique and exceptional care must to be taken when interpreting results. The collection of hundreds of scattering profiles in a short period of time runs the risk of producing hundreds of worthless datasets. Well-devised automated pipelines that cater for both the “one-off” novice and the highly experienced SAXS-User, incorporate data-quality checks so as to safeguard against over-zealous data interpretation. The delivery of such information in near real time using automated methods is not only useful for maintaining data-quality but also for providing an opportunity to refocus an investigation toward where it is often required—to the quality of the sample placed into an X-ray beam.

In the present paper we described the major steps of SAXS data collection and analysis, which have already been automated and can therefore be performed in a high-throughput mode. The automated interpretation at present includes the calculation of the overall structural parameters and *ab initio* low resolution shape determination. More advanced approaches like hybrid analysis (e.g., rigid body modelling in terms of high resolution or homology structures [51]), determination of oligomeric compositions of mixtures [77] or structural characterization of flexible macromolecules (ensemble analysis [19]) require additional information about the system under study. These analyses are still performed semiautomatically and need user input for running the relevant programs and for assessing the results. There are however developments towards automation of the more advanced analysis of the SAXS data [45, 78] and one can expect that high-throughput mode will become available for these applications in the near future.

Acknowledgements

We thank Dr David Jacques of the MRC Laboratory of Molecular Biology, Cambridge, for measuring the SAXS data presented in Figures 4 and 5. This work was supported by the Bundesministerium für Bildung und Forschung (BMBF) project BIOSCAT [05K12YE1].

References

1. Blanchet CE, Svergun DI (2013) Small-angle X-ray scattering on biological macromolecules and nanocomposites in solution. *Annu Rev Phys Chem* 64:37–54
2. Jacques DA, Trehwella J (2010) Small-angle scattering for structural biology—expanding the frontier while avoiding the pitfalls. *Protein Sci* 19(4):642–657
3. Rambo RP, Tainer JA (2013) Accurate assessment of mass, models and resolution by small-angle scattering. *Nature* 496(7446):477–481
4. Mertens HD, Svergun DI (2010) Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *J Struct Biol* 172(1):128–141

5. Rambo RP, Tainer JA (2013) Super-resolution in solution X-ray scattering and its applications to structural systems biology. *Annu Rev Biophys* 42:415–441
6. Petoukhov MV, Svergun DI (2013) Applications of small-angle X-ray scattering to biomacromolecular solutions. *Int J Biochem Cell Biol* 45(2):429–437
7. Blobel J, Bernado P, Svergun DI et al (2009) Low-resolution structures of transient protein-protein complexes using small-angle X-ray scattering. *J Am Chem Soc* 131(12):4378–4386
8. Williamson TE, Craig BA, Kondrashkina E et al (2008) Analysis of self-associating proteins by singular value decomposition of solution scattering data. *Biophys J* 94(12):4906–4923
9. Cho HS, Schotte F, Dashdorj N et al (2013) Probing anisotropic structure changes in proteins with picosecond time-resolved small-angle X-ray scattering. *J Phys Chem B* 117(49):15825–15832
10. Roessle M, Manakova E, Laure I et al (2000) Time-resolved small angle scattering: kinetics and structural data from proteins in solution. *J Appl Cryst* 33(1):548–551
11. Vestergaard B, Groenning M, Roessle M et al (2007) A helical structural nucleus is the primary elongating unit of insulin amyloid fibrils. *PLoS Biol* 5(5):e134
12. Graewert MA, Svergun DI (2013) Impact and progress in small and wide angle X-ray scattering (SAXS and WAXS). *Curr Opin Struct Biol* 23(5):748–754
13. Grant TD, Luft JR, Wolfley JR et al (2011) Small angle X-ray scattering as a complementary tool for high-throughput structural studies. *Biopolymers* 95(8):517–530
14. Pertea M, Salzberg SL (2010) Between a chicken and a grape: estimating the number of human genes. *Genome Biol* 11(5):206
15. Uversky VN (2010) The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome. *J Biomed Biotechnol* 2010:568068
16. Uversky VN, Dunker AK (2010) Understanding protein non-folding. *Biochim Biophys Acta* 1804(6):1231–1264
17. Chouard T (2011) Structural biology: breaking the protein rules. *Nature* 471(7337):151–153
18. Bernado P, Svergun DI (2012) Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol Biosyst* 8(1):151–167
19. Bernado P, Mylonas E, Petoukhov MV et al (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. *J Am Chem Soc* 129(17):5656–5664
20. Webb B, Lasker K, Velazquez-Muriel J et al (2014) Modeling of proteins and their assemblies with the Integrative Modeling Platform. *Methods Mol Biol* 1091:277–295
21. Schneidman-Duhovny D, Kim SJ, Sali A (2012) Integrative structural modeling with small angle X-ray scattering profiles. *BMC Struct Biol* 12:17
22. Forster F, Webb B, Krukenberg KA et al (2008) Integration of small-angle X-ray scattering data into structural modeling of proteins and their assemblies. *J Mol Biol* 382(4):1089–1106
23. Hammel M (2012) Validation of macromolecular flexibility in solution by small-angle X-ray scattering (SAXS). *Eur Biophys J* 41(10):789–799
24. Rambo RP, Tainer JA (2010) Bridging the solution divide: comprehensive structural analyses of dynamic RNA, DNA, and protein assemblies by small-angle X-ray scattering. *Curr Opin Struct Biol* 20(1):128–137
25. Pelikan M, Hura GL, Hammel M (2009) Structure and flexibility within proteins as identified through small angle X-ray scattering. *Gen Physiol Biophys* 28(2):174–189
26. Guinier A, Fournet G (1955) Small angle scattering of X-rays. Wiley, New York
27. Glatter O, Kratky O (1982) Small angle X-ray scattering. Academic Press, London
28. Feigin LA, Svergun DI (1987) Structure analysis by small-angle x-ray and neutron scattering. Plenum, New York
29. Svergun DI, Koch MHJ, Timmins PA, May RP (2013) Small angle X-ray and neutron scattering from solutions of biological macromolecules, 1st edn. Oxford University Press, New York
30. Koch MH, Vachette P, Svergun DI (2003) Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. *Q Rev Biophys* 36(2):147–227
31. Jeffries CMJ, Trehwella J (2012) Small angle-scattering. In: Wall ME (ed) Quantitative biology: from molecular to cellular systems. CRC, Boca Raton, pp 113–152
32. Putnam CD, Hammel M, Hura GL et al (2007) X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q Rev Biophys* 40(3):191–285
33. Broennimann C, Eikenberry EF, Henrich B et al (2006) The PILATUS 1 M detector. *J Synchrotron Radiation* 13:120–130
34. Orthaber D, Bergmann A, Glatter O (2000) SAXS experiments on absolute scale with Kratky

- systems using water as a secondary standard. *J Appl Cryst* 33:218–225
35. Guinier A (1939) La diffraction des rayons X aux tres petits angles; application a l'etude de phenomenes ultramicroscopiques. *Ann Phys (Paris)* 12:161–237
 36. Mylonas E, Svergun DI (2007) Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering. *J Appl Cryst* 40:s245–s249
 37. Svergun DI (1992) Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J Appl Cryst* 25:495–503
 38. Svergun DI, Semenyuk AV, Feigin LA (1988) Small-angle scattering data treatment by the regularization method. *Acta Cryst A* 44:244–250
 39. Semenyuk AV, Svergun DI (1991) GNOM—a program package for small-angle scattering data processing. *J Appl Cryst* 24:537–540
 40. Glatter O (1977) A new method for the evaluation of small-angle scattering data. *J Appl Cryst* 10:415–421
 41. Bergmann A, Fritz G, Glatter O (2000) Solving the generalized indirect Fourier transformation (GIFT) by Boltzmann simplex simulated annealing (BSSA). *J Appl Cryst* 33(5):1212–1216
 42. Jeffries CM, Lu Y, Hynson RM et al (2011) Human cardiac myosin binding protein C: structural flexibility within an extended modular architecture. *J Mol Biol* 414(5):735–748
 43. Porod G (1982) General theory. In: Glatter O, Kratky O (eds) *Small-angle X-ray scattering*. Academic, London, pp 17–51
 44. Fischer H, de Oliveira M, Napolitano HB et al (2010) The molecular weight of proteins in solution can be determined from a single SAXS measurement on a relative scale. *J Appl Cryst* 43:101–109
 45. Petoukhov MV, Franke D, Shkumatov AV et al (2012) New developments in the ATSAS program package for small-angle scattering data analysis. *J Appl Cryst* 45(2):342–350
 46. Jacques DA, Guss JM, Svergun DI et al (2012) Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution. *Acta Cryst D* 68(6):620–626
 47. Stuhmann H (1970) New method for determination of surface form and internal structure of dissolved globular proteins from small-angle X-ray measurements. *Z Phys Chem Frankfurt* 72:177–182
 48. Stuhmann HB (1970) Interpretation of small-angle scattering functions of dilute solutions and gases. A representation of the structures related to a one-particle-scattering function. *Acta Cryst A* 26:297–306
 49. Svergun DI (1999) Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys J* 76(6):2879–2886
 50. Svergun DI, Barberato C, Koch MHJ (1995) CRY SOL—a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J Appl Cryst* 28:768–773
 51. Petoukhov MV, Svergun DI (2005) Global rigid body modelling of macromolecular complexes against small-angle scattering data. *Biophys J* 89(2):1237–1250
 52. Konarev PV, Petoukhov MV, Svergun DI (2001) MASSHA—a graphic system for rigid body modelling of macromolecular complexes against solution scattering data. *J Appl Cryst* 34:527–532
 53. Franke D, Svergun DI (2009) DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *J Appl Cryst* 42:342–346
 54. Petoukhov MV, Svergun DI (2007) Analysis of X-ray and neutron scattering from biomacromolecular solutions. *Curr Opin Struct Biol* 17(5):562–571
 55. Chacon P, Moran F, Diaz JF et al (1998) Low-resolution structures of proteins in solution retrieved from X-ray scattering with a genetic algorithm. *Biophys J* 74(6):2760–2775
 56. Chacon P, Diaz JF, Moran F et al (2000) Reconstruction of protein form with X-ray solution scattering and a genetic algorithm. *J Mol Biol* 299(5):1289–1302
 57. Svergun DI, Petoukhov MV, Koch MHJ (2001) Determination of domain structure of proteins from X-ray solution scattering. *Biophys J* 80(6):2946–2953
 58. Svergun DI, Volkov VV, Kozin MB et al (1996) New developments in direct shape determination from small-angle scattering 2. Uniqueness. *Acta Cryst A* 52:419–426
 59. Vigil D, Gallagher SC, Trewheila J et al (2001) Functional dynamics of the hydrophobic cleft in the N-domain of calmodulin. *Biophys J* 80(5):2082–2092
 60. Walther D, Cohen FE, Doniach S (2000) Reconstruction of low-resolution three-dimensional density maps from one-dimensional small-angle X-ray solution scattering data for biomolecules. *J Appl Cryst* 33:350–363
 61. Volkov VV, Svergun DI (2003) Uniqueness of ab initio shape determination in small angle scattering. *J Appl Cryst* 36:860–864
 62. Kozin MB, Svergun DI (2001) Automated matching of high- and low-resolution structural models. *J Appl Cryst* 34:33–41
 63. Franke D, Kikhney AG, Svergun DI (2012) Automated acquisition and analysis of small

- angle X-ray scattering data. *Nucl Instrum Meth Phys Res Sect A* 689:52–59
64. Blanchet CE, Zozulya AV, Kikhney AG et al (2012) Instrumental setup for high-throughput small- and wide-angle solution scattering at the X33 beamline of EMBL Hamburg. *J Appl Cryst* 45(3):489–495
 65. Classen S, Rodic I, Holton J et al (2010) Software for the high-throughput collection of SAXS data using an enhanced Blu-Ice/DCS control system. *J Synchrotron Radiat* 17(6):774–781
 66. David G, Perez J (2009) Combined sampler robot and high-performance liquid chromatography: a fully automated system for biological small-angle X-ray scattering experiments at the Synchrotron SOLEIL SWING beamline. *J Appl Cryst* 42(5):892–900
 67. Classen S, Hura GL, Holton JM et al (2013) Implementation and performance of SIBYLS: a dual endstation small-angle X-ray scattering and macromolecular crystallography beamline at the Advanced Light Source. *J Appl Cryst* 46(Pt 1):1–13
 68. Hura GL, Menon AL, Hammel M et al (2009) Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). *Nat Methods* 6(8):606–612
 69. Pernot P, Theveneau P, Giraud T et al (2010) New beamline dedicated to solution scattering from biological macromolecules at the ESRF. *J Phys Conf Ser* 247(1):012009
 70. Nielsen SS, Moller M, Gillilan RE (2012) High-throughput biological small-angle X-ray scattering with a robotically loaded capillary cell. *J Appl Cryst* 45(2):213–223
 71. Toft KN, Vestergaard B, Nielsen SS et al (2008) High-throughput Small Angle X-ray Scattering from proteins in solution using a microfluidic front-end. *Anal Chem* 80(10):3648–3654
 72. Schneidman-Duhovny D, Hammel M, Sali A (2010) FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res* 38:W540–W544
 73. Petoukhov MV, Konarev PV, Kikhney AG et al (2007) ATSAS 2.1—towards automated and web-supported small-angle scattering data analysis. *J Appl Cryst* 40(s1):s223–s228
 74. Nyam-Osor M, Soloviov DV, Kovalev YS et al (2012) Silver behenate and silver stearate powders for calibration of SAS instruments. *J Phys Conf Ser* 351(1):012024
 75. Konarev PV, Petoukhov MV, Volkov VV et al (2006) ATSAS 2.1, a program package for small-angle scattering data analysis. *J Appl Cryst* 39:277–286
 76. Rolbin Y, Kayushina RL, Feigin LA et al (1973) Calculation of the small-angle X-ray scattering intensity on a computer using a macromolecule model. *Kristallografiya* 18:701–705
 77. Konarev PV, Volkov VV, Sokolova AV et al (2003) PRIMUS - a Windows-PC based system for small-angle scattering data analysis. *J Appl Cryst* 36:1277–1282
 78. Wassenaar T, Dijk M, Loureiro-Ferreira N et al (2012) WeNMR: structural biology on the grid. *J Grid Comput* 10(4):743–767

Chapter 16

Automated Structure Determination from NMR Spectra

Elena Schmidt and Peter Güntert

Abstract

Three-dimensional structures of proteins in solution can be calculated on the basis of conformational restraints derived from NMR measurements. This chapter gives an overview of the computational methods for NMR protein structure analysis highlighting recent automated methods for the assignment of NMR spectra, the collection of conformational restraints, and the structure calculation.

Key words Protein structure, NMR structure determination, Automated assignment, Resonance assignment, NOESY assignment, Conformational restraints, Network anchoring, Constraint combination, Torsion angle dynamics, CYANA, FLYA

1 Introduction

Until some years ago NMR protein structure determination was a laborious undertaking that occupied a trained spectroscopist over several months for each new protein structure. It was then recognized that many of the time-consuming manual steps carried out by an expert during the process of spectral analysis could be accomplished by automated, computational approaches [1]. Today automated methods for NMR structure determination are playing an ever more prominent role and are superseding the conventional manual approaches to solving three-dimensional protein structures in solution. This chapter gives an introduction to automated NMR assignment and structure calculation methods. Parts of this chapter were first published in the doctoral thesis of Elena Schmidt [2] and in [3, 4].

In most cases, protein structure determination is performed by a standard sequence of steps that are illustrated in Fig. 1. In the following, this standard procedure [5–8] is described with emphasis on peak picking, chemical shift assignment, nuclear Overhauser effect (NOE) assignment, and structure calculation.

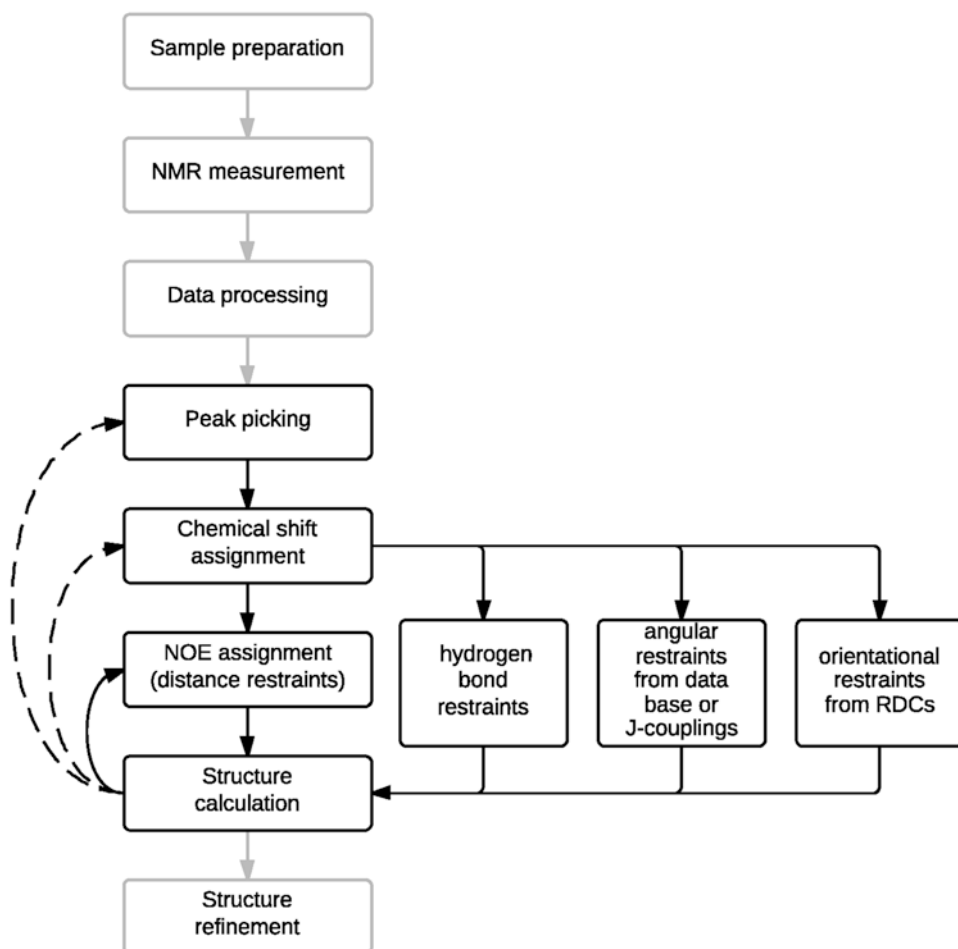


Fig. 1 Standard steps of NMR protein structure determination. The steps which are described in detail and can be applied iteratively are shown in *black boxes*. The utilization of structural information for the improvement of peak picking and chemical shift assignment is indicated with *dashed lines*

The first step in the process of protein structure determination is the preparation of the protein sample [9]. The protein to be studied is in most cases overexpressed in a bacterial system, which is usually grown on a $^{13}\text{C}/^{15}\text{N}$ isotopically enriched minimal medium. The protein is purified in order to obtain a sample of a few hundred microliters with a concentration in the (sub)millimolar range (>0.05 mM is sufficient using certain techniques).

The second step is the measurement of the atom signals with an NMR spectrometer, in which the protein sample is exposed to a strong magnetic field and a sequence of radiofrequency pulses. A set of different NMR experiments that differ from one another with respect to the pulse sequence are performed with the same sample. They result in experiment-specific signals that reveal the

covalent and spatial connectivities of the protein atoms. The third protein structure determination step is data processing. In order to obtain the NMR spectra, the measured time domain data is converted to frequency domain data using Fourier transformation and other techniques. In the fourth step “real” signals, which result from protein atoms, must be identified in the spectra and distinguished from noise and artifacts, which is referred to as peak picking.

The resulting peak lists are the basis for the next step, chemical shift assignment. The chemical shift values that are observed in the spectra are assigned to the corresponding protein atoms, since the relationship of the measured signals and the protein atoms is not known from the beginning.

The sixth step is NOE assignment. The cross peaks in NOESY spectra, which hold information about atom–atom distances in the 3D structure of the protein, are assigned to the respective atoms based on the chemical shift assignment. Distance restraints are deduced from the volumes of these peaks. In the seventh step the 3D structure is calculated based on distance restraints. The distance information can be complemented with angular restraints from chemical shifts or J-couplings, orientational restraints from RDCs, and hydrogen bond restraints. As soon as a preliminary structure is obtained, the structural information is used to improve the NOE assignment. This is done in several cycles.

Peak picking, chemical shift assignment, and NOE assignment are error-prone methods, especially when done with automated procedures. One possibility to reduce errors in automated procedures and to improve the structural quality is to apply these steps iteratively and to incorporate the structural information obtained from structure determination into peak picking and chemical shift assignment. This can be done by comparison of simulated spectra with real spectra in case of peak picking and structure-based chemical shift prediction and expected peak prediction in case of chemical shift assignment. Finally, the last structure determination step is to refine the structure using force fields adapted from molecular dynamics simulation packages.

2 Peak Picking

Peak picking is the procedure of extracting the positions of “real” peaks that result from molecule atoms from NMR spectra usually with several attributes like volume and shape. Resulting peak lists provide the basis for successful automated chemical shift and NOE assignment. Like chemical shift assignment, peak picking is a critical step in automated structure determination, since it is very prone to errors. Various automated programs exist for this task, but it is still common to pick peak lists manually or to refine them with manual intervention. Popular programs that can be used for peak

picking are CAPP [10], GIFA [11], AUTOPSY [12], ATNOS [13], SPARKY [14], NMRVIEW [15, 16], AURELIA [17], CcpNmr Analysis [18], and XEASY [19].

The challenge of peak picking is to identify all peaks even in overlapped regions and to distinguish between real peaks that are atom signals and noise or artifacts. The quality of the resulting peak lists has a large impact on chemical shift assignment, since real peaks that have not been included into the peak list can make it impossible to assign the respective atoms. On the other hand, additional peaks in the peak list may be confused with real peaks and can therefore lead to erroneous assignments.

Automated methods use several criteria to identify the set of real peaks in a spectrum. The most straightforward criteria are local maxima and the peak intensity. To identify also low-intensity peaks, the peak shape is considered. Simple shape attributes like line width are used, but also more advanced methods are applied to compare measured line shapes to ideal line shapes. Such procedures are implemented in AUTOPSY, CAPP, or ATNOS. Apart from just taking attributes of a specific peak into consideration, information about the experiment and, if available, about the assignment of the atoms and the protein structure can be included, e.g., in ATNOS. Peaks observed in other experiments or the symmetric properties of some spectra can help to distinguish between real peaks and artifacts, by providing the information whether or not a peak is expected at a specific position. This information can also be obtained from chemical shift assignments and the protein structure, which makes it possible to use peak picking, chemical shift assignment, and structure determination iteratively to refine a protein structure after the initial structure calculation.

3 Resonance Assignment

Every measured atom in a macromolecule has a specific chemical shift value, which depends on the chemical environment of this nucleus. The problem is that it is unknown from the start which atom leads to which chemical shift value. Revealing the relationship between atoms and chemical shifts is denoted as chemical shift assignment. Chemical shift assignment is necessary not only to evaluate the distance information in NOESY spectra for standard protein structure determination by NMR, but also in all cases in which atom-specific information has to be obtained from an NMR experiment. Examples are molecular interaction studies or alternative approaches for protein structure determination that are based on chemical shifts or RDCs, and investigations of protein dynamics.

To enable chemical shift assignment, several NMR experiments have to be performed that complement each other such that the connectivity of the atoms in a protein is represented. Based on the

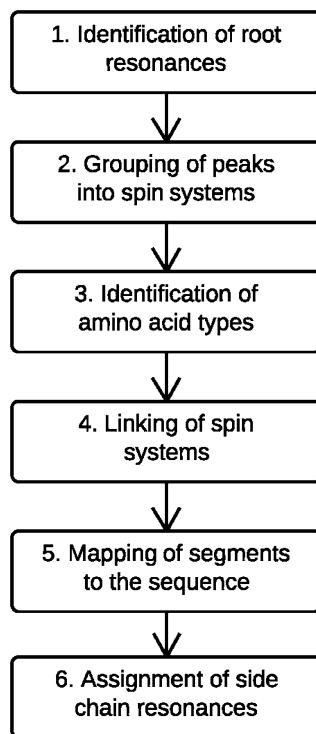


Fig. 2 Common steps of chemical shift assignment

knowledge about the covalent structure that can be deduced from the protein sequence, it is possible to establish the relationship between chemical shifts and atoms. Usually, a set of standard experiments are used to reveal the covalent atom connectivities.

Since the general strategy for chemical shift assignment has been described in the 1980s [20], there have been many attempts to establish an automated procedure for this process [1, 6, 21]. The published programs are based on various different optimization techniques including exhaustive search, best first approaches, genetic algorithms, and Monte Carlo methods. Some programs incorporate the whole chemical shift assignment process shown in Fig. 2, starting with peak lists or NMR spectra as input data and ending with the complete assignment of backbone and side chain atoms. Others are specialized in single steps of the process, usually sequence-specific assignment, in which the spin systems are assigned to the correct positions in the amino acid sequence of the protein. A selection of different programs for automated chemical shift assignment is listed in Table 1.

To date, none of these automatic approaches has been able to completely replace the procedure of manually assigning chemical shifts, since the completeness and reliability of the assignment results is often too low or the programs are only tested with

Table 1
Selected automated assignment programs

Program	Side chain	Spin systems	Input peak lists and notes
FLYA [39]	Yes	Yes	Any
GARANT [30, 31]	Yes	Yes	Standard 2D/3D through-bond, NOESY
PINE [24]	Yes	Yes	Standard 2D/3D through-bond, NOESY for RNA
AUTOASSIGN [22]	No	Yes	Standard 2D/3D through-bond
Moseley <i>et al.</i> [67]	No	Yes	3 3D/4D solid state
Li Sanctuary [68, 69]	Yes	Yes	Any general data set
ADAPT-NMR [25]	No	Yes	Integrated data collection
MARS [28]	No	No	Grouped chemical shifts
MATCH [26]	No	Yes	Designed for APSY [27]; possible with 3D
PASTA [70]	No	Yes	Any 2D/3D
PACES [71]	No	Assisted	Any triple resonance; user-assisted
MAPPER [72]	No	No	Shifts for amino acid stretches
DYNASSIGN [42]	Yes	Yes	Any
ASCAN [73]	Only	No	NOESY; backbone shifts
MONTE [74]	Yes	No	Grouped chemical shifts
TATAPRO [75]	No	Yes	Standard through-bond
Llinas <i>et al.</i> [76]	Yes	Yes	NOESY
Lukin <i>et al.</i> [77]	No	Yes	Standard through-bond
MCASSIGN2 [78]	Yes	Yes	2D/3D solid state; needs typing

The column “Side chain” gives information about what kind of assignment is done. “Yes” refers to assignment of backbone and side chains, “no” refers to assignment of the backbone, and “only” refers to assignment of the side chains. The column “Spin systems” gives information about whether the grouping of peaks into spin systems is done by the program. “yes” means that spin systems are generated automatically, “no” means that the peaks have to be grouped by the user, and “assisted” means that the grouping is done by the program but requires manual intervention

simulated data sets and fail when they are applied to real data of standard quality. The reason is that the human strategies that are used for manual assignment are difficult to convert into algorithms. Human experts follow the scheme for chemical shift assignment depicted in Fig. 2, but verify their decisions made in previous steps or complete intermediate results during the whole process, using

the information that they gained in the current step. It might be impossible, for example, to group some peaks correctly into two spin systems based solely on the information in the respective spectrum in cases in which the root resonances overlap. Later in the process these ambiguities can be resolved as soon as neighboring residues are sequence-specifically assigned.

Often, programs for automated assignment follow the standard steps for chemical shift assignment, but do not improve intermediate results that were obtained in previous steps. This can lead to erroneous assignments since not all information available was used. The same problem holds true if an automated procedure only performs single steps of the assignment process. The remaining steps have to be done manually by the user. Especially spin system identification, which is sometimes omitted by automated procedures, is time-consuming if done accurately and can in some cases only be completed during later steps of the manual assignment process. This is not possible if the results obtained in one step are not used to improve previous results. The situation is similar for peak picking. Peak lists are often updated during the assignment process, when chemical shift assignments can help to distinguish between noise or artifacts and real peaks. Automated assignment programs often use peak lists as input. Hence, the results produced by these programs strongly depend on the quality of these peak lists.

Another reason why programs for automated chemical shift assignment are not frequently used is that they are often restricted to a specific application. Many programs just allow for input from standard through-bond triple resonance spectra. In this case information from NOESY experiments, which is useful for side chain assignment, cannot be used and they cannot be applied whenever the measured data is different from the typical pattern.

In the next sections, a number of programs for automated chemical shift assignment are presented in more detail as examples for different optimization strategies and applications. AUTOASSIGN is the most popular program for automated chemical shift assignment of backbone atoms. PINE is able to perform all chemical shift assignment steps given standard peak lists obtained from through-bond spectra as input. ADAPT-NMR does not assign chemical shifts based on peak lists or NMR spectra, but directly controls the measurement of the data. MATCH is specialized on the assignment of high-dimensional APSY spectra. MARS can incorporate RDCs into the assignment process. The algorithm implemented in GARANT provided the bases for the FLYA automated assignment procedure, since it is not restricted to specific input spectra while providing complete chemical shift assignment and good results.

3.1 AUTOASSIGN

AUTOASSIGN [22] performs backbone assignment (HN , H^α , C , C^α , N , and C^β) based on a best-first approach using peak lists from the $[\text{}^{15}\text{N}, \text{}^1\text{H}]$ -HSQC experiment and standard 3D triple-resonance

experiments that are commonly used for manual backbone assignment.

The program starts with filtering of peak lists based on the N-H dimensions that are common in all lists and aligning the lists according to the resonances of isolated peaks. Root resonances for the grouping of peaks into spin systems are obtained from all peaks in the $^{15}\text{N}, ^1\text{H}$ -HSQC spectrum, and completed by HNCO peaks for resonances that were not included in the $^{15}\text{N}, ^1\text{H}$ -HSQC.

In the next step the peaks from the other lists are assigned to these generic spin system roots if their N and H frequencies match those of the root peak. The peaks of a generic spin system are grouped into two lists of chemical shifts, the C^α and the C' ladder, containing the chemical shifts of the respective amino acid itself and the previous one. According to the completeness, intensity, and degeneracy, the generic spin systems are categorized into distinct, overlapped, and weak. Based on the general chemical shift statistics of the BMRB and using Bayesian statistics for the C^α and C^β chemical shifts, possible amino acid type lists are created for the C^α and C' ladders of the generic spin systems. From these two lists a number of possible positions in the sequence of the respective dipeptide are deduced.

The following steps of linking different generic spin systems and assigning them to a position in the sequence are done using a constraint propagation method [23]. The principle is that these steps are first done with the category of distinct spin systems, followed by the overlapped and finally weak ones. The assignment of spin systems in the first category reduces the remaining assignment possibilities for all other spin systems and thereby simplifies assignment decisions for these spin systems. One cycle works as follows. All C^α and C' ladders are compared and a list of possible nearest neighbor links between the spin systems is created. The list is first ranked by the number of matching frequencies, and then by a match value. Links between neighbors are established if they build a unique one-to-one match within the group of the same number of matching frequencies (best-first approach). In the next step these spin system neighbor pairs are assigned to the sequence if there is a unique one-to-one match.

3.2 PINE

PINE (Probabilistic Interaction Network of Evidence algorithm) [24] performs probabilistic chemical shift assignment of backbone and side chain atoms and determines the secondary structure of the respective protein. It uses the protein sequence and two- and three-dimensional peak lists obtained from through-bond experiments as input. The method can in principle be extended to other experiments. Prior information about atom assignments can be included in the calculation.

The first step is to build up a network connecting the measured chemical shifts to all labeling possibilities, i.e., the atom names in the protein. This network is built up as follows. In order to group measured peaks into spin systems, similarity scores between peaks are determined based on the distances of peaks in common dimensions.

Starting with peaks of the most sensitive experiments, normally [^{15}N , ^1H]-HSQC or HNCO, spin systems are initialized and peaks with similarity scores greater than zero are added with a distance-dependent probability. Resulting spin systems cover peaks of spin system i and $i-1$. Connectivity scores between spin systems are calculated following the same scheme that was used for the calculation of similarity scores. According to the connectivity scores, the spin systems are grouped into triplet spin systems, since the usage of these triplet spin systems for assignment instead of single spin systems reduces the complexity of the network.

Amino acid typing is done by calculating a score for each combination of spin system and amino acid triplet that can be assigned to each other. The respective scoring of a single atom takes into account the BMRB chemical shift statistics and the secondary structure prediction with respect to the atom in the sequence. The score for a triplet is obtained by calculating the product of the respective atom scores in the triplet spin system.

The assignment is done in several iterations. The backbone assignment probabilities in the network are determined based on the amino acid scoring, connectivity experiments, backbone assignment in the previous iteration, and outlier detection. The topology and the probabilities of this network are changed during several iterations until a quasi-stationary state is reached, which means that topology and probabilities do not vary significantly. In each iteration an energy function is evaluated, and a belief propagation algorithm is applied to obtain an updated network and thereby probabilistic assignments as well as the probabilities for the secondary structure. Rather than a single assignment for each atom, several probability-weighted possibilities are obtained.

After the backbone assignment a separate network model for each amino acid is generated and the belief propagation algorithm is applied to obtain probabilistic side chain assignments.

3.3 ADAPT-NMR

ADAPT-NMR (Assignment-directed Data collection Algorithm utilizing a Probabilistic Toolkit in NMR) [25] provides fully automated backbone assignment and secondary structure prediction. It implements the concept of iterative chemical shift assignment and data measurement using a probabilistic network approach.

The algorithm starts with the generation of probabilistic spin systems on the basis of the [^{15}N , ^1H]-HSQC spectrum. A probabilistic spin system has different attributes. Each attribute may have

different assignments at the same time. For example, attributes are the measured chemical shifts.

During the optimization cycle, evaluation of the probabilistic spin systems leads to the experiment type and the plane to be recorded next. The respective peaks are picked automatically and a probability based on peak and experiment characteristics is determined using several machine learning techniques. Subsequently, a pseudoenergy model is used to update probabilistic spin systems.

For high-probability peaks in 3D spectra that do not match the [$^{15}\text{N}, ^1\text{H}$]-HSQC and overlapping regions additional spin systems can be introduced during the calculation. If the spin system quality is below a specified threshold, it is optimized in a new cycle of data collection and probabilistic network update. When the threshold is reached, the assignment step is done.

The core part of the assignment step originates from the PINE algorithm with some modifications, e.g., a fully probabilistic implementation. In the assignment step probabilities for chemical shift assignments, secondary structure states, and outlier chemical shift values are determined. As soon as an initial assignment is available, the assignment and the secondary structure are also considered for the data collection, i.e., selection of the next experiment and plane. At this point data is explicitly collected for spin systems that are weakly linked in order to maximize the information gain in the data collection step.

3.4 MATCH

MATCH (Mimetic Algorithm and Combinatorial Optimization Heuristics) [26] provides chemical shift assignment of the protein backbone using peak lists as input. The program was implemented for the use with APSY [27] spectra, but can also be used with standard triple experiments. It combines an evolutionary algorithm with a local optimization routine.

During the initialization process, the measured frequencies are grouped into spin systems and these are assembled into bigger fragments of a given maximum size. Therefore, connectivities between spin systems are identified based on a scoring function. Each spin system obtains the list of all possible fragments it is part of and finally, isolated spin systems are removed and control parameters for the calculation are set according to the degree of ambiguity of the data.

To start with, an initial population of assignment solutions is generated. Each assignment solution is obtained by randomly selecting a fragment of maximum length and mapping it to the position in the sequence with the highest sequence-specific score. The sequence-specific score evaluates the agreement of the measured chemical shifts with the chemical shifts statistics of the BMRB. The process is repeated first with the remaining fragments of the same length. Afterwards the length of the fragments to be mapped is decremented. Fragments including spin systems that are

already mapped are not considered in further selections. During the local optimization, pairs of fragments that could be mapped to the sequence but have not yet been assigned permanently are selected randomly and tested for compatibility of spin systems and matching adjacent spin systems and possibly they are interchanged. If the sequence-specific score exceeds a given threshold, a temporary assignment is created, which means that the assignment will not be changed by the local optimization anymore. Permanent assignments are established if a specific assignment can be found in more than a given fraction of the population.

In the global optimization routine the solutions of several individuals are combined into new assignment solutions. The individuals are ranked according to their sequence-specific score and the best-scored solutions are used to build up the next generation. The optimization ends when all atoms have permanently been assigned.

If the calculation does not converge, the whole process is repeated with modified control parameters. The optimization is repeated several times to obtain independent results. An assignment is output if it is present in at least 50 % of the independent runs.

3.5 MARS

The program MARS [28, 29] performs backbone chemical shift assignment using intra- and inter-residue chemical shifts, which have to be grouped into “pseudoresidues” by the user. The program can include RDCs into the assignment process if a 3D protein structure is available.

The first step of the algorithm is the generation of all possible sequential connections between pseudoresidues. Connections that do not agree with the experimental data are removed at later stages. Distances between the experimental chemical shifts of the pseudoresidues and chemical shift predictions for possible sequence assignments are calculated. These predictions are obtained based on the standard BMRB statistics, correcting for neighbor residue effects and secondary structure effects. According to these distances, the sequence positions are ranked and a pseudoenergy is determined.

The optimization starts with a random assignment of the pseudoresidues to positions in the amino acid sequence. Starting from a random pseudoresidue, segments of five residues including this residue are assembled based on the connectivity information. The segments are mapped onto all possible positions of the sequence. The probability for an assignment is calculated using the pseudoenergy and the solutions are ranked, the minimum representing the best solution. If starting with the last pseudoresidue of the respective segments leads to the same assignment solution, the solution is considered reliable and a penalty for all other possibilities is included into the pseudoenergy function. The procedure is repeated with all pseudoresidues as starting points for the assembly. In subsequent steps the segment size is decreased. The procedure

is further repeated adding varying noise to the predicted chemical shifts. Consistent solutions are considered as reliable. If RDC measurements and a 3D protein structure are available for the respective protein, the agreement between the measured RDCs and RDCs calculated from the provided protein structure is included into the distance function.

3.6 GARANT

GARANT (General Algorithm for Resonance AssignmeNT) [30] is a program for automatic chemical shift assignment of backbone and side chain atoms. It can be used for chemical shift assignment solely based on a set of peak lists and the protein amino acid sequence, but can also assign chemical shifts based on a given 3D protein structure [31].

The general concept of GARANT is that the connectivities between the atoms of a protein that can be revealed with different NMR experiments are represented by a network of expected peaks and protein atoms. The mapping of this network to the network of measured peaks and their chemical shifts leads to an assignment of the atoms to the respective chemical shifts. The optimization problem of finding the mapping possibility that corresponds to the correct assignment solution is solved using an evolutionary algorithm in combination with a local optimization routine.

Both optimization routines use a scoring scheme based on the concept of mutual information. The different terms of the scoring function for the evolutionary optimization evaluate the assignment of a chemical shift to an atom, based on the agreement with general chemical shift statistics, the mapping of an expected peak to a measured peak, and the agreement of the chemical shifts of single signals and the respective atom chemical shift. Ambiguous peak mappings lead to a reduced score. The global score results from the sum over the terms resulting from all atom and peak mappings.

The evolutionary optimization routine is controlled by a simulated annealing temperature schedule. It uses a population of 50 assignment solutions by default. For the construction of a new generation, parent solutions with a high global score are favored. Mappings for a specific residue are adapted from parent solutions as far as possible. If no parent solution is available, those of residues with similar spin systems are considered. If no mapping solution could be determined that way, new mappings are generated.

The local optimization routine selects unmapped or ambiguously mapped expected peaks and evaluates the assignment of neighboring atoms based on a local score and reassigns these atoms, if necessary. The local optimization routine also uses the mutual information-based scoring scheme, except that for the scoring of an atom only contributions that are directly related to the atom are included.

An advantage of the GARANT algorithm compared to many other methods is that in principle it can solve the assignment

problem using every combination of spectra that contain sufficient information for the assignment. Even though the algorithm is generally applicable to any kind of common NMR spectrum, the specification of a given set of spectra is done within the program and can only be extended by changes of the C++ source code.

The fully automated chemical shift assignment program GARANT was introduced in 1996. In the respective publications [30, 31] the application of GARANT was demonstrated by assigning different proteins up to a sequence length of 165 amino acids, based on data sets that consisted solely of homonuclear 2D spectra as well as data sets consisting of 3D spectra. GARANT has been used in various other projects, and has also been adapted for calculations with APSY [27] spectra and applied to the assignment of 4–7D APSY spectra in different applications [32, 33].

To obtain full automation of chemical shift assignment starting from NMR spectra, GARANT was combined with the automated peak picking program AUTOPSY [12] and a program for calibration and filtering, PICS [34]. GARANT was combined with AUTOPSY, the structure calculation program CYANA, and the molecular dynamics simulation package OPALp to achieve fully automated structure determination, including peak picking, chemical shift assignment, NOE assignment, structure calculation, and energy refinement [35]. It has been shown that this strategy is in principle also applicable to sparse data [36, 37]. Even structure determination based solely on NOESY data was successful [38].

3.7 FLYA

The new FLYA automated chemical shift assignment procedure has been implemented and applied to several targets [4, 39]. As described above, various programs for automated chemical shift assignment have been developed before, but none of these approaches has become a standard procedure. The main reasons for this are the generally low accuracy of the assignment results and restrictions to specific applications.

The aforementioned GARANT program for automated chemical shift assignment is based on an optimization strategy that can in principle be applied to every kind of NMR spectrum and provides good assignment results compared to other programs. However, assignment calculations with GARANT take relatively long, i.e., several hours in some cases, the application to nonstandard data sets is not straightforward and the accuracy of the results leaves room for improvement. This situation led to the development of a new algorithm, FLYA, with the following objectives: (1) improving the accuracy of the chemical shift assignment, (2) improving the flexibility of the method in order to address a wider range of problems, (3) shortening the run time of the algorithm, and (4) incorporating automated resonance assignment into the CYANA software in order to simplify the application of the algorithm in conjunction with other CYANA modules, e.g., NOE distance restraint assignment and structure calculation by torsion angle dynamics.

The resulting implementation of the FLYA automated assignment algorithm [39] in the CYANA software package includes a modified expected peak generation procedure, a new scoring scheme, and various further improvements of the optimization algorithm over the earlier GARANT approach.

NMR resonance assignment is based on experiments that correlate nuclear spins such that they give rise to cross peaks in multidimensional spectra. Assignment experiments are chosen to complement each other in such a way that the connectivity of the atoms in a protein can be represented by a network of peaks that are expected to be observed. Mapping this network of expected peaks with unknown positions to the unassigned measured peaks with known positions provides an assignment of the frequencies to the spins [30, 31]. The FLYA algorithm for automated backbone and side chain resonance assignment uses this general approach to assign all kinds of NMR spectra. It is implemented in the software package CYANA [40, 41]. As input, FLYA uses exclusively the sequence of the protein and unassigned peak lists from any combination of multidimensional solution-state or solid-state NMR spectra.

All experimental data is used simultaneously in order to exploit optimally the redundancy present in the input peak lists and to avoid potential pitfalls of assignment strategies in which results obtained in a given step remain fixed input data for subsequent steps. Instead of prescribing a specific assignment strategy, the FLYA resonance assignment algorithm generates the peaks expected in a given spectrum by applying a set of rules for through-bond or through-space polarization transfer, and determines the resonance assignment by constructing an optimal mapping between the expected peaks, assigned by definition but having unknown positions, and the measured peaks, initially unassigned but with known positions in the spectrum [30, 31, 39, 42].

The rules for generating expected peaks have been implemented for many different solution-state and solid-state NMR experiments. Expected peaks for experiments like NOESY or DARR, which give signals between atoms that are close in space, are obtained using random structures of the respective proteins [39]. An expected peak is generated for each atom pair up to a given cutoff on the maximal distance between the two atoms in the ensemble of random structures. This will generate expected peaks only if the atoms are close together in the primary structure, e.g., for intra-residual and sequential distances. It corresponds to the generation of expected peaks for NOE-based experiments in solution NMR. Expected peaks for all other experiments are obtained based on the covalent connections between atoms. For each experiment the covalent bond patterns that hold this information are provided to the algorithm in the CYANA library file. It is straightforward to add new experiments or to modify the rules for existing experiments.

The best mapping of expected peaks to measured peaks is obtained using an evolutionary optimization routine that works with a population of individuals, each representing an assignment solution for the protein. This evolutionary optimization is complemented by local optimization. Solutions that are produced during the optimization are generated such that the search space of an expected peak for a mapping is defined by a chemical shift statistics (by default from the BMRB [43], or user defined), the deviations of the measured frequencies of measured peaks that are assigned to the same atom remain within a given tolerance, and an expected peak can be mapped to only one measured peak. The first generation of solutions is generated randomly, but subject to these conditions. In each generation a local optimization algorithm takes small parts of a mapping back and reassigns the expected peaks for a defined number of iterations, 15,000 is default. Afterwards the different solutions of one generation are recombined into a new generation. The individuals and the specific parts of an individual that contribute to a new individual are selected via a scoring function. The solution that maximizes this function is given as the final assignment at the end of the calculation.

The global score for complete assignment solutions evaluates four attributes of an assignment solution, the distribution of chemical shift values with respect to the given shift statistics, the alignment of peaks assigned to the same atom, the completeness of the assignment, and a penalty for chemical shift degeneracy. The global score G is defined by

$$G = \frac{\sum_{a \in A} \left[W_1(a) Q_1(a) + \sum_{n \in N'_a} W_2(a, n) Q_2(a, n) / b(n) \right]}{\sum_{a \in A_0} \left[W_1(a) + \sum_{n \in N_a} W_2(a, n) \right]}$$

The term A_0 denotes the set of all atoms for which expected peaks exist, $A \subseteq A_0$ the set of assigned atoms, N_a the set of expected peaks for atom a , and $N'_a \subseteq N_a$ the subset of expected peaks that are mapped to a measured peak. $b(n)$ refers to the ambiguity of the assignment and equals the number of expected peaks that are assigned to the same measured peak as expected peak n . Unassigned atoms and unmapped peaks contribute through the normalization by the denominator. Relative weights of the individual contributions are given by $w_1(a)$ and $w_2(a, n)$ and in [39] these were set to $w_1(a)=4$ and $w_2(a, n)=1$ for all calculations. The quality measure $Q_1(a)$ represents the agreement of the average chemical shift $\bar{\omega}(a)$ in the chemical shift list of atom a with the corresponding general chemical shift statistics. Similarly, $Q_2(a, n)$ measures the agreement between the chemical shift $\omega(a, n)$ of atom a obtained from the measured peak to which the expected peak n is mapped and the

average frequency of the atom in the assigned peaks of the corresponding spectrum [39]. The quality measures Q are designed such that a perfect match corresponds to $Q=1$, $Q<1$ in all other cases, a deviation that is considered “as bad as no assignment” yields $Q=0$, and an infinitely large deviation $Q=-\infty$. Consequently, the global score G is normalized such that $G=1$ for a (hypothetical) perfect assignment, and $G<1$ in all other cases.

The main difference to solution NMR lies in the rules for generating expected peaks, which have been implemented for many different solid-state NMR experiments (Table 1).

To improve and assess the accuracy of the assignment, m independent runs of the algorithm are performed with different random seeds. For each atom a consensus chemical shift is computed from the values obtained in the individual runs [34, 35, 39]. The consensus chemical shift $\tilde{\omega}(a)$ for an atom a is the value that maximizes the function

$$\mu(\omega) = \frac{1}{m} \sum_{j=1}^m \exp \left(-\frac{1}{2} \left(\frac{\omega - \omega_j(a)}{\epsilon(a)} \right)^2 \right),$$

where $\omega_j(a)$ is the chemical shift value obtained for atom a in run j , and $\epsilon(a)$ is the chemical shift tolerance. The maximum value of this function, $\mu(\tilde{\omega}(a))$, is a measure of the self-consistency of the chemical shift values obtained in the individual runs of the algorithm, since it approximately equals the fraction of runs that yielded a chemical shift value within the tolerance $\epsilon(a)$ from the consensus value $\tilde{\omega}(a)$. This quantity can be calculated without knowledge of reference assignments. If all chemical shift values are identical, then $\mu(\tilde{\omega}(a))=1$. We consider assignments with $\mu(\tilde{\omega}(a)) \geq 0.8$ as “strong” or self-consistent, all others as “weak.” Weak assignments should be considered as tentative, although they are correct in many cases.

In the following, an example of a CYANA macro for a standard chemical shift assignment is shown, which refers to the published assignment calculations [39]. Executing this macro leads to a standard chemical shift assignment of the protein atoms and evaluation of the assignment results based on reference shifts. NOE assignment and structure calculation are not done.

```
1. assigns_accH:=0.03
2. assigns_accC:=0.4
3. assigns_accN:=0.4
4. assignpeaks:=N15NOESY,C13NOESY,\
5. C13HSQC,N15HSQC,HCCHTOCSY,HCCHCOSY,HNCA,HNca
   CO,HNCO,HNcoCA,\
6. CBCANH,CBCAcoNH,HBHAcoNH,CcoNH,HCcoNH
```



```
7. cyanalib
8. read seq protein.seq
9. flya shiftreference=protein.prot runs=20 assignpeaks=
   $assignpeaks
```

Lines 1–3: The tolerances are set for the assignment calculation and the comparison with the reference chemical shifts in the file “protein.prot,” 0.03 ppm for hydrogen atoms and 0.4 ppm for carbon and nitrogen atoms.

Line 4: The peak list names of NOESY experiments are specified. Expected peaks for all experiments are generated according to the entries in the CYANA library file. In case of expected peaks for NOESY experiments distances are deduced from a random structure (“start.pdb”). Alternatively, the user can specify a structure.

Line 5/6: The peak list names of the through-bond experiments are specified.

Line 7: The CYANA standard library is read.

Line 8: The protein amino acid sequence is read from the file “protein.seq”.

Line 9: The command flya, which is specified in the macro “flya.cya,” is executed. The chemical shifts in file “protein.prot” obtained from manual assignment are used as reference for the automated assignment. The number of independent assignment runs is set to 20. The content of the variable “assignpeaks” is given as input for the parameter “assignpeaks” of the macro “flya.cya”.

4 NOE Assignment

Structure determination with distance restraints obtained from NOESY spectra relies on the fact that ^1H nuclei which are separated by less than 5 Å in the protein lead to cross-peaks in the spectra according to the isolated two-spin approximation. The cross-peak volumes are proportional to the inverse sixth power of the distance between the nuclei. Since neighbor effects may weaken the observed signal, the presence of a cross-peak provides only an upper distance limit for the distance between the two hydrogen atoms [44]. In order to convert the peak volumes observed in the spectra to distance restraints, a spectrum-dependent calibration constant has to be determined.

During NOE assignment the observed cross-peaks in a NOESY spectrum are assigned to the corresponding atom pairs according to the resonance assignment, in order to generate a list of distance restraints, which provides the basis for structure calculation. An illustration of a protein structure with all distance restraints, which

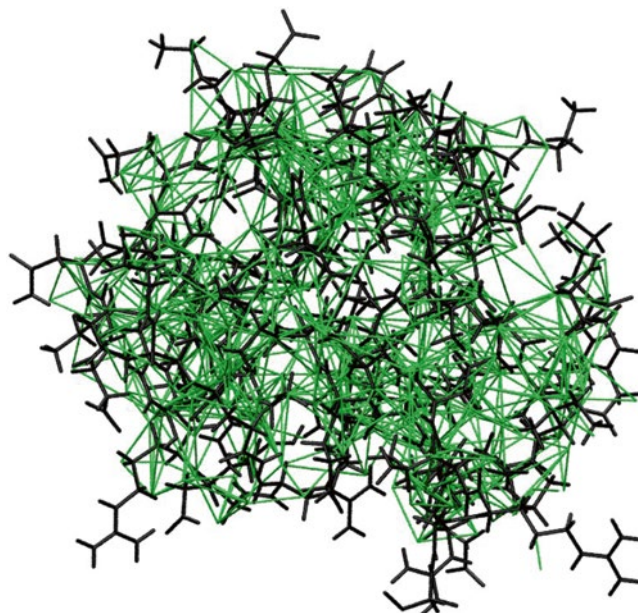


Fig. 3 Distance restraints that were obtained from ^{13}C and ^{15}N -resolved NOESY spectra and used as input for the structure calculation of the protein. The protein atoms are shown in *black*. Distance restraints in *green* connect corresponding hydrogen atoms

were obtained from NOESY spectra and used for the structure calculation, is shown in Fig. 3.

Obtaining a comprehensive set of distance restraints from a NOESY spectrum is in practice by no means straightforward. Resonance and peak overlap turn NOE assignment into an iterative process in which preliminary structures, calculated from limited numbers of distance restraints, serve to reduce the ambiguity of the cross peak assignments. Additional difficulties may arise from spectral artifacts and noise, and from the absence of expected signals because of fast relaxation. These inevitable shortcomings of NMR data collection are the main reason why laborious interactive procedures have dominated this central step of NMR protein structure determination for a long time. Automated procedures follow the same general scheme as the interactive approach but do not require manual intervention during the assignment/structure calculation cycles. Two main obstacles have to be overcome by an automated method starting without any prior knowledge of the structure: First, the number of cross peaks with unique assignment based on chemical shift alignment alone is in general not sufficient to define the fold of the protein [7]. An automated method must therefore have the capability to use also NOESY cross peaks that cannot (yet) be assigned unambiguously. Second, the automated program must be able to cope with the erroneously picked

or inaccurately positioned peaks and with the incompleteness of the chemical shift assignment of typical experimental data sets. An automated procedure needs devices to substitute for the intuitive decisions made by an experienced spectroscopist in dealing with the imperfections of experimental NMR data.

Besides semiautomatic approaches [45–47], several algorithms have been developed for the automated analysis of NOESY spectra given the chemical shift assignments of the backbone and side chain resonances, namely NOAH [48, 49], ARIA [50–53], AUTOSTRUCTURE [54], KNOWNOE [55], CANDID [56] and a similar algorithm implemented in CYANA [57], PASD [58], and a Bayesian approach [59]. Automated NOE assignment algorithms generally require a high degree of completeness of the backbone and side chain chemical shift assignments [60].

4.1 Combined Automated NOE Assignment and Structure Calculation with CYANA

A widely used algorithm for the automated interpretation of NOESY spectra is implemented in the NMR structure calculation program CYANA [41, 57]. This algorithm is a re-implementation of the former CANDID algorithm [56] on the basis of a probabilistic treatment of the NOE assignment, combined in an iterative process that comprises seven cycles of automated NOE assignment and structure calculation, followed by a final structure calculation using only unambiguously assigned distance restraints. Between subsequent cycles, information is transferred exclusively through the intermediary 3D structures. The molecular structure obtained in a given cycle is used to guide the NOE assignments in the following cycle. Otherwise, the same input data are used for all cycles, that is, the amino acid sequence of the protein, one or several chemical shift lists from the sequence-specific resonance assignment, and one or several lists containing the positions and volumes of cross peaks in 2D, 3D, or 4D NOESY spectra. The input may further include previously assigned NOE upper distance bounds or other previously assigned conformational restraints for the structure calculation.

In each cycle, first all assignment possibilities of a peak are generated on the basis of the chemical shift values that match the peak position within given tolerance values, and the quality of the fit is expressed by a Gaussian probability, P_{shifts} . Second, in all but the first cycle the probability $P_{\text{structure}}$ for agreement with the preliminary structure from the preceding cycle, represented by a bundle of conformers, is computed as the fraction of the conformers in which the corresponding distance is shorter than the upper distance bound plus the acceptable distance restraint violation cutoff. The precision of the structure determination normally improves with each subsequent cycle. Accordingly, the cutoff for acceptable distance restraint violations in the calculation of $P_{\text{structure}}$ is tightened from cycle to cycle. Third, each assignment possibility is evaluated for its network anchoring (see below), which is quantified by

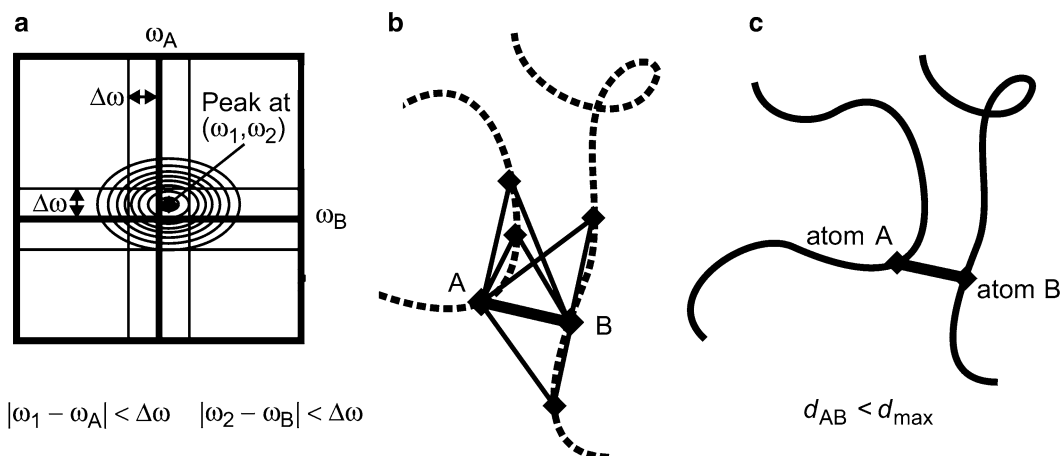


Fig. 4 Three conditions that must be fulfilled by a valid assignment of a NOESY cross peak to two protons A and B in the CYANA automated NOESY assignment algorithm: **(a)** Agreement between the proton chemical shifts ω_A and ω_B and the peak position (ω_1, ω_2) within a tolerance of $\Delta\omega$. **(b)** Spatial proximity in a (preliminary) structure. **(c)** Network anchoring. The NOE between protons A and B must be part of a network of other NOEs or covalently restricted distances that connect the protons A and B indirectly through other protons

the probability P_{network} . Only assignment possibilities for which the product of the three probabilities is above a threshold,

$$P_{\text{tot}} = P_{\text{shifts}} \cdot P_{\text{structure}} \cdot P_{\text{network}} \geq P_{\text{min}},$$

are accepted (Fig. 4). Cross peaks with a single accepted assignment yield a conventional unambiguous distance restraint. Otherwise, an ambiguous distance restraint is generated that embodies multiple accepted assignments.

4.2 Ambiguous Distance Restraints

Because of the limited accuracy of chemical shift values and peak positions many NOESY cross peaks cannot be attributed to a single unique spin pair but have an ambiguous NOE assignment comprising multiple spin pairs. Ambiguous distance restraints [61] provide a powerful concept for handling ambiguities in the initial, chemical shift-based NOESY cross peak assignments. Prior to the introduction of ambiguous distance restraints in the ARIA algorithm [53], in general only unambiguously assigned NOEs could be used as distance restraints in the structure calculation. Since the majority of NOEs cannot be assigned unambiguously from chemical shift information alone, this lack of a general way to include ambiguous data into the structure calculation considerably hampered the performance of early automatic NOESY assignment algorithms. When using ambiguous distance restraints, every NOESY cross peak is treated as the superposition of the signals from each of its possible assignments by applying relative weights proportional to the inverse sixth power of the corresponding

interatomic distances. A NOESY cross peak with a unique assignment possibility gives rise to an upper bound b on the distance $d(\alpha, \beta)$ between two hydrogen atoms, α and β . A NOESY cross peak with $n > 1$ assignment possibilities can be interpreted as the superposition of n degenerate signals and interpreted as an ambiguous distance restraint, $d_{\text{eff}} \leq b$, with the “effective” or “ r^{-6} -summed” distance

$$d_{\text{eff}} = \left(\sum_{k=1}^n d_k^{-6} \right)^{-1/6}.$$

Each of the distances $d_k = d(\alpha_k, \beta_k)$ in the sum corresponds to one assignment possibility to a pair of hydrogen atoms, α_k and β_k . The effective distance d_{eff} is always shorter than any of the individual distances d_k . Thus, an ambiguous distance restraint will be fulfilled by the correct structure provided that the correct assignment is included among its assignment possibilities, regardless of the possible presence of other, incorrect assignment possibilities. Ambiguous distance restraints make it possible to interpret NOESY cross peaks as correct conformational restraints also if a unique assignment cannot be determined at the outset of a structure determination. Including multiple assignment possibilities, some but not all of which may later turn out to be incorrect, does not result in a distorted structure but only in a decrease of the information content of the ambiguous distance restraints.

4.3 Network Anchoring

Each assignment possibility is evaluated for its network anchoring, i.e., its embedding in the network formed by the assignment possibilities of all the other peaks and the covalently restricted short-range distances. The network anchoring probability P_{network} that the distance corresponding to an assignment possibility is shorter than the upper distance bound plus the acceptable violation is computed given the assignments of the other peaks but independent from knowledge of the three-dimensional structure. Contributions to the network anchoring probability for a given, “current” assignment possibility result from other peaks with the same assignment, from pairs of peaks that connect indirectly the two atoms of the current assignment possibility via a third atom, and from peaks that connect an atom in the vicinity of the first atom of the current assignment with an atom in the vicinity of the second atom of the current assignment. Short-range distances that are constrained by the covalent geometry take, for network anchoring, the same role as an unambiguously assigned NOE. Individual contributions to the network anchoring of the current assignment possibility are expressed as probabilities, P_1, P_2, \dots , that the distance corresponding to the current assignment possibility satisfies the upper distance bound. The network anchoring probability is obtained from the

individual probabilities as $P_{\text{network}} = 1 - (1 - P_1) \cdot (1 - P_2) \cdots$, which is never smaller than the highest probability of an individual network anchoring contribution.

4.4 Constraint Combination

In practice, spurious distance restraints may arise from the misinterpretation of noise and spectral artifacts, in particular at the outset of a structure determination, before 3D structure-based filtering of the restraint assignments can be applied. The key technique used in CYANA to reduce structural distortions from erroneous distance restraints is “constraint combination” [56]. Ambiguous distance restraints are generated with combined assignments from different, in general unrelated, cross peaks (Fig. 5). The basic property of ambiguous distance restraints that the restraint will be fulfilled by the correct structure whenever at least one of its assignments is correct, regardless of the presence of additional, erroneous assignments, then implies that such combined restraints have a lower probability of being erroneous than the corresponding original restraints, provided that the fraction of erroneous original restraints is smaller than 50 %. Constraint combination aims at minimizing the impact of such imperfections on the resulting structure at the expense of a temporary loss of information. It is applied to medium- and long-range distance restraints in the first two cycles of combined automated NOE assignment and structure calculation with CYANA.

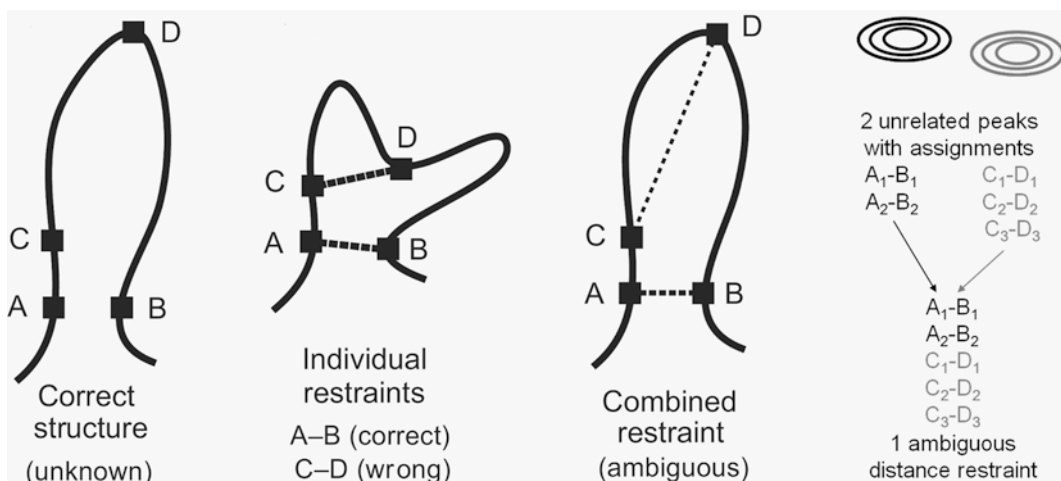


Fig. 5 Schematic illustration of the effect of constraint combination in the case of two distance restraints, a correct one connecting atoms A and B, and a wrong one between atoms C and D. A structure calculation that uses these two restraints as individual restraints that have to be satisfied simultaneously will, instead of finding the correct structure (shown, schematically, in the *first panel*), result in a distorted conformation (*second panel*), whereas a combined restraint that will be fulfilled already if one of the two distances is sufficiently short leads to an almost undistorted solution (*third panel*). The formation of a combined restraint from the assignments of two peaks is shown in the *right panel*

5 Structure Calculation

The three-dimensional protein structure is calculated using the list of distance restraints, which are obtained from NOESY spectra. Commonly used programs for structure calculation are CYANA [40] (formerly DYANA [41]/DIANA [62]), Xplor-NIH [63], and CNS [64] (formerly X-PLOR [65]), where CYANA is the program most widely used for NMR structure calculation [66].

The most efficient algorithm for the calculation of 3D protein structures from distance restraints, which is also implemented in CYANA, performs simulated annealing by molecular dynamics simulation in torsion angle space. The simulated annealing procedure minimizes a potential energy function, which takes distance restraints, angle restraints, and a repulsive potential into account. Atom distance information from NOESY spectra can be complemented, e.g., by angle restraints, orientational restraints, and hydrogen bond restraints.

A CYANA structure calculation with automated NOE assignment can be completed in less than 1 h for a 10–15 kDa protein, provided that the structure calculations can be performed in parallel, for instance on a Linux cluster system.

In the following, an example of a CYANA macro for a standard combined automated NOE assignment and structure calculation is shown. Executing this macro performs the automated NOE assignment and the structure calculation of a protein.

```
1. peaks := c13.peaks,n15.peaks,aro.peaks
2. prot := demo.prot
3. restraints := demo.aco
4. tolerance := 0.04, 0.03, 0.45
5. structures := 100,20
6. steps := 10000
7. randomseed := 434726
8. cyanalib
9. read seq protein.seq
10. noeassign peaks=$peaks prot=$prot autoaco
```

Line 1: The names of the input NOESY peak lists are specified.

Line 2: The names of the input chemical shift lists are specified. In this case, there is one chemical shift list that is used for all peak lists.

Line 3: The names of additional input restraint files, in this case a file with torsion angle restraints, are specified.

Line 4: Tolerances are set for the NOE assignment calculation, i.e., 0.04/0.03 ppm for hydrogen atoms in the indirect/direct dimensions, and 0.45 ppm for carbon and nitrogen atoms.

Line 5: The numbers of conformers that are calculated (100) and analyzed [20] are specified.

Line 6: The number of torsion angle dynamics steps in the structure calculation is specified.

Line 7: The random number generator seed for generating initial structures is specified.

Line 8: The CYANA standard library is read.

Line 9: The protein amino acid sequence is read from the file “protein.seq”.

Line 10: The command noeassign, which is specified in the macro “noeassign.cya” is executed with the given NOESY peak lists and chemical shift list(s) as input. The option “autoaco” specifies that weak torsion angle restraints for the Ramachandran plot and staggered side chain rotamers will be generated and used for the structure calculations.

References

1. Moseley HNB, Montelione GT (1999) Automated analysis of NMR assignments and structures for proteins. *Curr Opin Struct Biol* 9:635–642
2. Schmidt E (2014) Institute of biophysical chemistry. Goethe University, Frankfurt am Main
3. Güntert P (2011) In: Lian LY, Roberts GCK (eds). *Protein NMR spectroscopy: principal techniques and applications*. Wiley, Chichester, UK. pp 159–192
4. Schmidt E, Gath J, Habenstein B et al (2013) Automated solid-state NMR resonance assignment of protein microcrystals and amyloids. *J Biomol NMR* 56:243–254
5. Billeter M, Wagner G, Wüthrich K (2008) Solution NMR structure determination of proteins revisited. *J Biomol NMR* 42:155–158
6. Gronwald W, Kalbitzer HR (2004) Automated structure determination of proteins by NMR spectroscopy. *Prog Nucl Magn Reson Spectrosc* 44:33–96
7. Güntert P (2003) Automated NMR protein structure calculation. *Prog Nucl Magn Reson Spectrosc* 43:105–125
8. Malmödin D, Billeter M (2005) High-throughput analysis of protein NMR spectra. *Prog Nucl Magn Reson Spectrosc* 46:109–129
9. Muskett FW (2011) In: Lian LY, Roberts GCK (eds). *Protein NMR spectroscopy: practical techniques and applications*. Wiley, Chichester, UK. pp 5–21
10. Garrett DS, Powers R, Gronenborn AM, Clore GM (1991) A common sense approach to peak picking two-, three- and four-dimensional spectra using automatic computer analysis of contour diagrams. *J Magn Reson* 95:214–220
11. Pons JL, Malliavin TE, Delsuc MA (1996) Gifa V. 4: a complete package for NMR data set processing. *J Biomol NMR* 8:445–452
12. Koradi R, Billeter M, Engeli M et al (1998) Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *J Magn Reson* 135:288–297
13. Herrmann T, Güntert P, Wüthrich K (2002) Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *J Biomol NMR* 24:171–189
14. Goddard TD, Kneller DG (2001) *Sparky 3*. University of California, San Francisco
15. Johnson BA (2004) Using NMRView to visualize and analyze the NMR spectra of macromolecules. *Methods Mol Biol* 278:313–352
16. Johnson BA, Blevins RA (1994) NMR view—a computer program for the visualization and analysis of NMR data. *J Biomol NMR* 4:603–614
17. Neidig KP, Geyer M, Gorler A et al (1995) Aurelia, a program for computer-aided analysis of multidimensional NMR spectra. *J Biomol NMR* 6:255–270
18. Vranken WF, Boucher W, Stevens TJ et al (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins* 59:687–696

19. Bartels C, Xia TH, Billeter M et al (1995) The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J Biomol NMR* 6:1–10
20. Wüthrich K, Wider G, Wagner G, Braun W (1982) Sequential resonance assignments as a basis for determination of spatial protein structures by high-resolution proton nuclear magnetic resonance. *J Mol Biol* 155:311–319
21. Guerry P, Herrmann T (2011) Advances in automated NMR protein structure determination. *Q Rev Biophys* 44:257–309
22. Zimmerman DE, Kulikowski CA, Huang YP et al (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. *J Mol Biol* 269:592–610
23. Zimmerman D, Kulikowski C, Wang LZ et al (1994) Automated sequencing of amino acid spin systems in proteins using multidimensional HCC(CO)NH-TOCSY spectroscopy and constraint propagation methods from artificial intelligence. *J Biomol NMR* 4:241–256
24. Bahrami A, Assadi AH, Markley JL, Eghbalnia HR (2009) Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. *PLoS Comput Biol* 5:e1000307
25. Bahrami A, Tonelli M, Sahu SC, Singarapu KK et al (2012) Robust, integrated computational control of NMR experiments to achieve optimal assignment by ADAPT-NMR. *PLoS One* 7:e33173
26. Volk J, Herrmann T, Wüthrich K (2008) Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH. *J Biomol NMR* 41:127–138
27. Hiller S, Fiorito F, Wüthrich K, Wider G (2005) Automated projection spectroscopy (APSY). *Proc Natl Acad Sci U S A* 102:10876–10881
28. Jung YS, Zweckstetter M (2004) Mars—robust automatic backbone assignment of proteins. *J Biomol NMR* 30:11–23
29. Jung YS, Zweckstetter M (2004) Backbone assignment of proteins with known structure using residual dipolar couplings. *J Biomol NMR* 30:25–35
30. Bartels C, Güntert P, Billeter M, Wüthrich K (1997) GARANT—a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *J Comput Chem* 18:139–149
31. Bartels C, Billeter M, Güntert P, Wüthrich K (1996) Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. *J Biomol NMR* 7:207–213
32. Fiorito F, Hiller S, Wider G, Wüthrich K (2006) Automated resonance assignment of proteins: 6D APSY-NMR. *J Biomol NMR* 35:27–37
33. Hiller S, Wider G, Wüthrich K (2008) APSY-NMR with proteins: practical aspects and backbone assignment. *J Biomol NMR* 42:179–195
34. Malmodin D, Papavoine CHM, Billeter M (2003) Fully automated sequence-specific resonance assignments of heteronuclear protein spectra. *J Biomol NMR* 27:69–79
35. López-Méndez B, Güntert P (2006) Automated protein structure determination from NMR spectra. *J Am Chem Soc* 128:13112–13122
36. Ikeya T, Takeda M, Yoshida H et al (2009) Automated NMR structure determination of stereo-array isotope labeled ubiquitin from minimal sets of spectra using the SAIL-FLYA system. *J Biomol NMR* 44:261–272
37. Scott A, López-Méndez B, Güntert P (2006) Fully automated structure determinations of the Fes SH2 domain using different sets of NMR spectra. *Magn Reson Chem* 44:S83–S88
38. Ikeya T, Jee J-G, Shigemitsu Y et al (2011) Exclusively NOESY-based automated NMR assignment and structure determination of proteins. *J Biomol NMR* 50:137–146
39. Schmidt E, Güntert P (2012) A new algorithm for reliable and general NMR resonance assignment. *J Am Chem Soc* 134:12817–12829
40. Güntert P (2009) Automated structure determination from NMR spectra. *Eur Biophys J* 38:129–143
41. Güntert P, Mumenthaler C, Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J Mol Biol* 273:283–298
42. Schmucki R, Yokoyama S, Güntert P (2009) Automated assignment of NMR chemical shifts using peak-particle dynamics simulation with the DYNASSIGN algorithm. *J Biomol NMR* 43:97–109
43. Ulrich EL, Akutsu H, Doreleijers JF et al (2008) BioMagResBank. *Nucleic Acids Res* 36:D402–D408
44. Levitt MH (2008) Spin dynamics: basics of nuclear magnetic resonance. Wiley, New York
45. Duggan BM, Legge GB, Dyson HJ, Wright PE (2001) SANE (structure assisted NOE evaluation): an automated model-based approach for NOE assignment. *J Biomol NMR* 19:321–329
46. Güntert P, Berndt KD, Wüthrich K (1993) The program ASNO for computer-supported collection of NOE upper distance constraints

- as input for protein structure determination. *J Biomol NMR* 3:601–606
47. Meadow RP, Olejniczak ET, Fesik SW (1994) A computer-based protocol for semiautomated assignments and 3D structure determination of proteins. *J Biomol NMR* 4:79–96
 48. Mumenthaler C, Braun W (1995) Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry. *J Mol Biol* 254:465–480
 49. Mumenthaler C, Güntert P, Braun W, Wüthrich K (1997) Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. *J Biomol NMR* 10:351–362
 50. Habeck M, Rieping W, Linge JP, Nilges M (2004) NOE assignment with ARIA 2.0: the nuts and bolts. *Methods Mol Biol* 278:379–402
 51. Linge JP, Habeck M, Rieping W, Nilges M (2003) ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* 19:315–316
 52. Rieping W, Habeck M, Bardiaux B (2007) ARIA2: automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics* 23:381–382
 53. Nilges M, Macias MJ, ODonoghue SI, Oschkinat H (1997) Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from beta-spectrin. *J Mol Biol* 269:408–422
 54. Huang YJ, Tejero R, Powers R, Montelione GT (2006) A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins* 62:587–603
 55. Gronwald W, Moussa S, Elsner R et al (2002) Automated assignment of NOESY NMR spectra using a knowledge based method (KNOWNOE). *J Biomol NMR* 23:271–287
 56. Herrmann T, Güntert P, Wüthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J Mol Biol* 319:209–227
 57. Güntert P (2004) Automated NMR structure calculation with CYANA. *Methods Mol Biol* 278:353–378
 58. Kuszewski J, Schwieters CD, Garrett DS et al (2004) Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear overhauser enhancement spectra and chemical shift assignments. *J Am Chem Soc* 126:6258–6273
 59. Hung LH, Samudrala R (2006) An automated assignment-free Bayesian approach for accurately identifying proton contacts from NOESY data. *J Biomol NMR* 36:189–198
 60. Jee J, Güntert P (2003) Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. *J Struct Funct Genomics* 4:179–189
 61. Nilges M (1995) Calculation of protein structures with ambiguous distance restraints—automated assignment of ambiguous NOE crosspeaks and disulfide connectivities. *J Mol Biol* 245:645–660
 62. Güntert P, Braun W, Wüthrich K (1991) Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. *J Mol Biol* 217:517–530
 63. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. *J Magn Reson* 160:65–73
 64. Brünger AT, Adams PD, Clore GM (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr Sect D Biol Crystallogr* 54:905–921
 65. Brünger AT (1992) X-PLOR, Version 3.1. A system for X-ray crystallography and NMR. Yale University Press, New Haven
 66. Williamson MP, Craven CJ (2009) Automated protein structure calculation from NMR data. *J Biomol NMR* 43:131–143
 67. Moseley HNB, Sperling LJ, Rienstra CM (2010) Automated protein resonance assignments of magic angle spinning solid-state NMR spectra of beta 1 immunoglobulin binding domain of protein G (GB1). *J Biomol NMR* 48:123–128
 68. Li KB, Sanctuary BC (1997) Automated resonance assignment of proteins using heteronuclear 3D NMR. 2. Side chain and sequence-specific assignment. *J Chem Inf Comput Sci* 37:467–477
 69. Li KB, Sanctuary BC (1997) Automated resonance assignment of proteins using heteronuclear 3D NMR. 1. Backbone spin systems extraction and creation of polypeptides. *J Chem Inf Comput Sci* 37:359–366
 70. Leutner M, Gschwind RM, Liermann J et al (1998) Automated backbone assignment of labeled proteins using the threshold accepting algorithm. *J Biomol NMR* 11:31–43

71. Coggins BE, Zhou P (2003) PACES: Protein sequential assignment by computer-assisted exhaustive search. *J Biomol NMR* 26:93–111
72. Güntert P, Salzmann M, Braun D, Wüthrich K (2000) Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER. *J Biomol NMR* 18: 129–137
73. Fiorito F, Herrmann T, Damberger FF, Wüthrich K (2008) Automated amino acid side-chain NMR assignment of proteins using ^{13}C - and ^{15}N -resolved 3D [^1H , ^1H]-NOESY. *J Biomol NMR* 42:23–33
74. Hitchens TK, Lukin JA, Zhan YP et al (2003) MONTE: an automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins. *J Biomol NMR* 25:1–9
75. Atreya HS, Sahu SC, Chary KVR, Govil G (2000) A tracked approach for automated NMR assignments in proteins (TATAPRO). *J Biomol NMR* 17:125–136
76. Grishaev A, Llinás M (2002) Protein structure elucidation from NMR proton densities. *Proc Natl Acad Sci U S A* 99:6713–6718
77. Lukin JA, Gove AP, Talukdar SN, Ho C (1997) Automated probabilistic method for assigning backbone resonances of (C-13, N-15)-labeled proteins. *J Biomol NMR* 9: 151–166
78. Hu KN, Qiang W, Tycko R (2011) A general Monte Carlo/simulated annealing algorithm for resonance assignment in NMR of uniformly labeled biopolymers. *J Biomol NMR* 50:267–276

Solid-State Nuclear Magnetic Resonance Spectroscopy for Membrane Protein Structure Determination

Peter J. Judge, Garrick F. Taylor, Hugh R.W. Dannatt, and Anthony Watts

Abstract

Solid-state NMR (ssNMR) is a versatile technique that can provide high-resolution (sub-angstrom) structural data for integral membrane proteins embedded in native and model membrane environments. The methodologies for a priori structure determination have for the most part been developed using samples with crystalline and fibrous morphologies. However, the techniques are now being applied to large, polytopic membrane proteins including receptors, ion channels, and porins. ssNMR data may be used to annotate and refine existing structures in regions of the protein not fully resolved by crystallography (including ligand-binding sites and mobile solvent accessible loop regions). This review describes the spectroscopic experiments and data analysis methods (including assignment) used to generate high-resolution structural data for membrane proteins. We also consider the range of sample morphologies that are appropriate for study by this method.

Key words Solid-state nuclear magnetic resonance, Membrane proteins

1 Introduction

Membrane proteins enable organisms to functionalize and specialize the selectively permeable barriers that divide their cells and subcellular compartments. Although structural data are available for only a small number, the diversity and importance of these proteins are apparent from their relative abundance in the genome, typically comprising 20–40 % of all open reading frames [1, 2]. Membrane proteins are of major interest to the pharmaceutical industry, since receptors and ion channels alone account for at least half of all human drug targets [3]. Rational, structure-based drug design is dependent on a detailed knowledge of high-resolution protein structures, yet ligand-binding sites are often poorly resolved in crystallographic studies of receptors [4]. Membrane proteins pose considerable challenges to conventional diffraction methods, due to their unwillingness to form ordered crystals of sufficient size for

analysis (up to μm for nanofocus beamlines). Although wild-type membrane proteins may be studied using solution NMR, these studies are also limited by the necessary application of nonnative solubilizing detergents.

Solid-state NMR (ssNMR) is a highly versatile and rapidly developing technique uniquely able to provide high-resolution structures of integral membrane proteins embedded in native or native-like lipid environments [5]. For an up-to-date list of membrane proteins and peptides whose structures have been solved by ssNMR see www.drorlist.com/nmr/SPNMR.html. Whereas the applicability of solution NMR to insoluble macromolecules and large protein complexes is limited by the dependence of the spectral line widths on the molecular tumbling rate (τ_c^{-1}), theoretically protein size is not restrictive for ssNMR, excepting spectral overlap which may be (partially) relieved using selective isotope labeling (*see* Subheading 2) [6]. The molecular tumbling rate τ_c^{-1} approximates $0.75kT/\pi r^3\eta$ where r is the molecular radius and η is the viscosity.

The electrical field surrounding an atom nucleus within an organic molecule is generally asymmetrical, since covalent bonds distort the electron distribution. The degree of shielding of a nucleus from the magnetic field by the electron cloud, which is observed in NMR spectra as the chemical shift, is therefore dependent on molecular orientation. In solution NMR, the chemical shift anisotropy is averaged by the rapid molecular tumbling, and spectra show narrow isotropic peaks. In ssNMR samples, molecular tumbling is either significantly slower or entirely absent, and so molecules are present in every possible orientation relative to the external magnetic field. This results in spectra with broad lines comprising all possible chemical shifts superimposed for each nucleus. In addition, the effects of dipolar coupling between nearby spins, which are dependent on the relative orientation of the coupled spins, are preserved in the solid state and can lead to very rapid signal decay via spin diffusion.

Although the information about orientation provided by dipolar couplings and the chemical shift anisotropy (CSA) are useful structural constraints, the removal of these effects provides increased spectral resolution and coherence lifetimes. Dipolar couplings and CSA have an angular dependence ($3\cos^2\theta - 1$) relative to the orientation of the external magnetic field, and so are eliminated at the so-called magic angle ($\theta = 54.74^\circ$). Magic angle spinning (MAS) ssNMR probes are designed to enable samples to be rapidly rotated at the magic angle in order to average the CSA and to reduce dipolar couplings, producing spectra that primarily consist of isotropic peaks (a side effect of MAS is the appearance of additional peaks known as spinning side bands at lower spinning speeds). Under MAS conditions, dipolar couplings between spins may be selectively reintroduced by the application of radiofrequency (RF)

pulses, allowing magnetization to be transferred between nuclei (permitting a “walk” along the polypeptide backbone, or down amino acid side chains to be made) or to measure homo- and heteronuclear dipolar couplings between nuclei from which internuclear distances can be readily calculated (*see* Subheading 4).

2 Sample Preparation and Labeling

Sample preparation is a major challenge for biological solid-state NMR, given the relatively high concentration ($\sim\mu\text{M}$) and amount (10–300 μl) of labeled material required for the experiments. Not all nuclei found in biological molecules are detectable by NMR. The only naturally abundant spin $\frac{1}{2}$ nucleus found extensively in proteins is ^1H (spin $\frac{1}{2}$ nuclei are easy to detect by NMR). Most proteins prepared for solid-state NMR experiments are isotopically enriched with ^{13}C or ^{15}N , both of which have low natural abundance and are spin $\frac{1}{2}$ nuclei. Since proteins produced for NMR analysis are typically expressed recombinantly, there is great flexibility in the isotope labeling strategies which may be used, since proteins may be either labeled on all atoms (uniformly) or selectively labeled using labeled metabolic precursors or amino acids.

Solution NMR studies of proteins center around the detection of ^1H signals, as ^1H has the largest gyromagnetic ratio of all stable nuclei, and therefore gives strong NMR signals. However in the solid state, at slow or moderate MAS rates (up to 30 kHz), the strong dipolar couplings between ^1H nuclei enable fast spin diffusion, producing extremely broad linewidths (typically tens of kHz), meaning that proton-detected experiments have not been widely used for biological molecules [7]. The perdeuteration of all non-exchangeable ^1H sites in conjunction with only partial restoration (approx. 10 %) of ^1H nuclei at exchangeable sites increases spectral resolution even at slow MAS rates, albeit at the cost of reduced sensitivity [8–10].

Recent technological advances have enabled sample rotors to be spun at rates of 40 kHz or higher, at which speeds the ^1H – ^1H homonuclear dipolar couplings are strongly reduced, such that the advantages of the high gyromagnetic ratio (γ) of ^1H nuclei may be exploited without the need for significant dilution of ^1H spins. The spectral linewidths of ^1H resonances are approximately inversely proportional to the sample spinning speed in these fast MAS experiments [7].

2.1 Labeling Strategies for ssNMR

Unless ^1H -detected ssNMR is to be used, protein and peptide samples must be produced which incorporate NMR isotope labels either in all atoms of a given type (uniform labeling) or at specific sites within the molecules. The labeling strategy chosen will be dependent on a number of factors, including the size of the

protein, the type of information that will be obtained from the sample, the time available for NMR experiments, the expression system or synthesis methodology used to produce the sample, and the cost of the labels themselves.

It is important to consider the size of the protein to be studied, since the number of resonances which will need to be assigned is directly related to the molecular weight. For instance a small antimicrobial peptide of 20 amino acids could easily be uniformly labeled without fear of spectral crowding which would make interpretation of spectra difficult. However uniform labeling of a seven-transmembrane receptor (300 amino acids) would result in very crowded spectra with a high degree of spectral overlap.

Other important factors to be considered are time and cost. Isotope-enriched media and compounds are expensive; how expensive depends on the isotope and labeling scheme desired. As a consequence of this it may not be possible to, for instance, iteratively label each amino acid type or combinations of amino acids to facilitate unambiguous assignment. The use of selectively labeled glycerol ($[1,3-^{13}\text{C}]$ and $[2-^{13}\text{C}]$ -glycerol) to grow recombinant bacteria results in the site-specific labeling of the amino acids produced by those bacteria. The specificity of the labeling depends on how those amino acids are produced. Amino acids derived from either the glycolytic or pentose phosphate pathways (alanine, cysteine, glycine, histidine, leucine, phenylalanine, serine, tryptophan, tyrosine, and valine) can be almost 100 % labeled with either ^{12}C or ^{13}C . The remaining amino acids are produced from precursors in the citric acid cycle and give rise to nonrandom mixtures of isotopomers [11].

The specific labeling of amino acids allows resonances to be assigned due to the unique cross-peaks that will occur between spins within certain residues. For instance valines will be unique in having strong $\text{C}\alpha\text{--C}\beta$ cross-peaks in the proton-driven spin diffusion (PDSD) spectrum acquired from protein produced with $[2-^{13}\text{C}]$ -glycerol. In addition the reduced amount of labeled spins vastly reduces overlap, allowing resonances to be identified which otherwise may have been obscured. The labeling scheme also reduces the number of adjacent carbon atoms, eliminating many of the J-couplings which, although not resolved in a typical solid-state NMR spectra, do contribute to the linewidth, limiting the available resolution. The dilution of ^{13}C nuclei reduces dipolar truncation effects and enables longer range contacts to be measured. Removal of some $^{13}\text{C}\text{--}^{13}\text{C}$ dipolar coupling also decreases linewidths and therefore improves resolution, further facilitating assignment. This labeling scheme has been successfully used before by Castellani et al. in the assignment of the α -spectrin Src homology 3 domain (SH3) [12] as well as directly on fibrils with the CA150.WW2 domain by Becker et al. [13].

2.2 Sample Morphology and Environment

A major advantage of solid-state NMR over other methods for membrane protein structure determination is the ability to study proteins in model and native membranes. The bilayer environment that surrounds membrane proteins in living organisms has a direct influence on the structure, dynamic, and function of membrane proteins and its complexity is difficult to replicate in biophysical experiments [14, 15]. Unless a protein naturally occurs in a given natural membrane at high concentration with few additional components (e.g., bacteriorhodopsin in the *H. salinarum* purple membrane) [16], it must be purified, concentrated, and reconstituted into a suitable model environment [17]. Care must be taken to ensure that the protein of interest is active in the chosen model membrane to ensure that the structure that is determined is biologically valid.

A common approach used in preparing membrane proteins for structure determination by MAS NMR is reconstitution into lipid vesicles, to produce proteoliposomes. In general the process requires the slow removal of detergent from a solubilized membrane sample, so that the surfactant concentration falls below the critical micelle concentration (CMC) [18]. In general multilamellar vesicles provide a good mimetic for cellular membranes and may be used to generate samples with a high protein content. Discussion of reconstitution methodologies is beyond the scope of this review; however detailed procedures have recently been published for reconstitution of two *Mycobacterium tuberculosis* proteins [19]. A recent exciting development for membrane protein sample preparation is the use of cell-free expression systems, which facilitate labeling of specific amino acids and reduce the complexity of protein purification procedures. The methodology has been successfully applied to the MscL ion channel, but is yet to be proven for more complex membrane proteins [20, 21].

Oriented sample NMR requires the magnetic or mechanical alignment of lipid bilayers relative to the magnetic field. Mechanical alignment generally requires the layering of membranes onto a solid substrate support (usually glass coverslips), although imperfections in orientation generally lead to spectral broadening and a decrease in the precision of the restraints in orientation generated from the NMR data. Increasingly, magnetically oriented discoidal membranes, including nanodiscs and bicelles, are being used for this type of NMR experiment [22] since they allow a higher sample filling factor within the RF coil of the NMR probe [23] but are able to include a wider variety of lipid types than conventional mechanically oriented samples [24]. One disadvantage of magnetically aligned discoidal membrane systems is the sensitivity of orientation to changes in temperature [25] and the presence of a protein may also lead to a change in the alignment of bicelles [26]. An alternative method of producing aligned samples by ultracentrifugation accompanied by simultaneous solvent evaporation is described by Gröbner et al. in 1997 [27].

3 Experimental Setup

3.1 Equipment

The basic requirements for ssNMR experiment are as follows:

1. Access to a solid-state NMR spectrometer with three channels (one each for ^1H , ^{13}C , ^{15}N), equipped with a triple-resonance MAS probe capable of taking ssNMR rotors of 4 mm diameter or below (Fig. 1). NMR spectrometer manufacturers include Bruker, Agilent Technologies, and Jeol.
2. A reliable source of compressed gas (up to 5 mbar) for sample spinning is required (boil-off gas from a liquid nitrogen tank is generally used). It may be appropriate to cool the sample below room temperature, in which case a gas cooler will also be required.
3. At least 1× zirconium oxide solid-state NMR rotor, 4 mm diameter or below (Fig. 2).
4. Reference samples, pre-packed, including KBr (used for setting the magic angle), adamantane (used for shimming and pulse power measurements), and $\text{U-}^{13}\text{C-}^{15}\text{N}$ glycine (used for pulse power measurements).



Fig. 1 Solid-state NMR probes. *Left:* Photograph of an MAS probe heads showing the brown stator unit and angled central hole, into which a zirconium oxide MAS rotor containing sample is inserted. The protective aluminum head is rotated to screen the rotor and the probe height is adjusted within the superconducting magnet to ensure that the sample experiences a uniform magnetic field along its length. Ducts within the stator unit channel compressed air around the rotor, allowing the sample to spin. *Right:* Photograph of a low-E field-oriented sample probe head, showing the rectangular hole to hold a stack of glass slides, onto which membranes have been deposited. The glass coverslips are typically held within a rectangular thin glass container

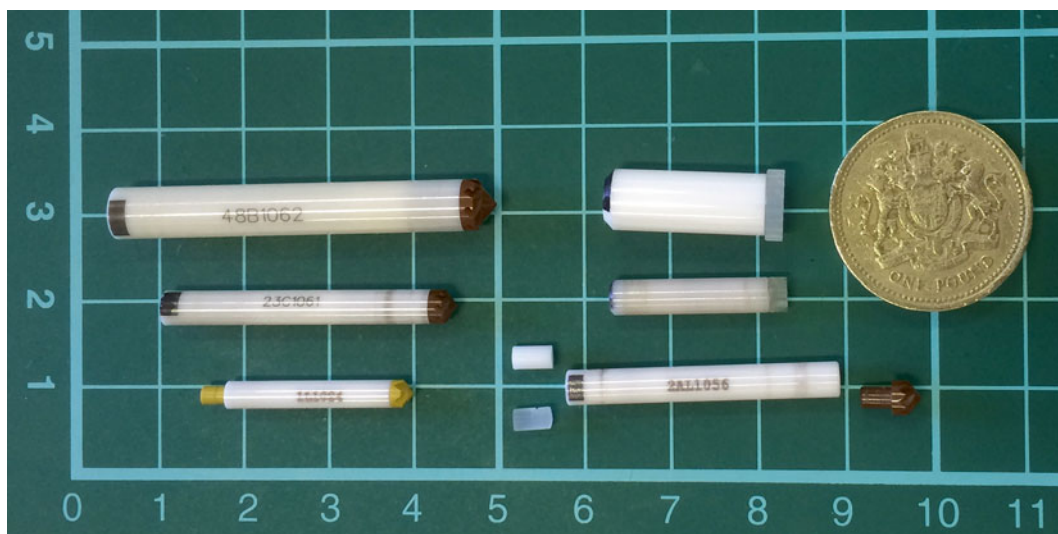


Fig. 2 A selection of different sized magic angle spinning (MAS) rotors manufactured by Bruker and Agilent Technologies. Samples are held in position by Kel-F or Teflon spacers (*center, bottom*) and the rotation is driven by compressed air passing across the vaned drive tip (*bottom right*). Additional rubber seals (not shown) may be used between the spacers, to maintain the hydration of protein and membrane samples. A British pound coin (*top right*) is given for size comparison

3.2 Sample Packing

Solid-state NMR samples which are to be studied under magic angle spinning conditions must be rapidly rotated at an angle of 54.7° to the magnetic field. Membrane samples are typically packed into a zirconium oxide rotor which is capable of withstanding the centrifugal forces generated by fast MAS (Fig. 1). The sample is held in place by Teflon or Kel-F spacers and it is important to ensure that the packed rotor is well balanced and that the sample is positioned at the correct place to ensure that it experiences a homogenous magnetic field [28].

The centrifugal forces that result from MAS can sediment the sample against the inner walls of the rotor, allowing buffer to be removed from the center of the rotor and replaced with more sample. This process can be repeated several times to ensure that the rotor is packed optimally with protein. One way to avoid the need for this process, as well as to facilitate the packing of the small-diameter rotors used for MAS frequencies of >30 kHz, is to use a rotor filling tool in which the sample is centrifuged directly into the rotor [29, 30]. Buffers for solid-state NMR studies of membrane proteins are generally chosen to closely model physiological conditions, with a pH range of 6.0–8.0 and a NaCl concentration of 50–150 mM [31].

4 ssNMR Experiments for Structure Determination

Strategies used by solid-state NMR spectroscopists to generate protein structures generally fall into one of the two different approaches depending on the number and precision of the distance or angle constraints generated by the data. Experiments used to determine the dihedral φ and ψ angles of amino acids within a protein are described first (Subheading 4.1). Subsequently experiments which are used to generate high-precision distance and constraints in orientation are outlined.

4.1 Protein Structures from Dihedral Angles

Local geometrical restraints are produced by comparing the measured chemical shifts of assigned atoms within a protein to average chemical shifts for the same atoms in disordered random coil protein regions. The general aim is to produce a structure calculated from numerous low-precision distance and angle measurements, rather than a few highly accurate measurements (*see* Subheading 4.4). Before these restraints can be generated, a combination of several experiment types is required in order to assign the chemical shifts to specific nuclei within a protein molecule.

In general the first experiments require ^{13}C – ^{13}C broadband recoupling to allow individual spin systems (nuclei belonging to an individual amino residue) to be identified and in many cases the process begins with the identification of the amino acid type (i.e., proline, threonine). The position of an amino acid in the protein primary sequence can be determined using heteronuclear correlation experiments that require the transfer of magnetization from nitrogen to carbon atoms, both intra- and inter-residue. Two-dimensional NCA and NCO experiments allow proximal residues to be correlated with each other and allow site-specific identification of residues when the primary sequence of the protein is known. In favorable cases of signal to noise and transfer efficiency, three-dimensional NCACX and NCOCX experiments can be used to further reduce ambiguity in the resonance assignments.

4.2 The Proton-Driven Spin Diffusion Experiment

The proton-driven spin diffusion (PDSD) is one of many broadband dipolar recoupling experiments, for example, dipolar-assisted rotational resonance (DARR) [32] and radio frequency-driven recoupling (RFDR) [33] which reintroduce the homonuclear dipolar couplings between low- γ nuclei (such as ^{13}C and ^{15}N) across a broad range of chemical shifts, an interaction which is lost under MAS conditions. During a PDSD experiment magnetization is moved via spin-diffusion between ^{13}C atoms along chemical bonds, resulting in a two-dimensional spectrum that is symmetrical about a diagonal line which connects the bottom left-hand corner to the top right corner. Correlations between ^{13}C nuclei are identified by off-diagonal cross-peaks. At short mixing times one-bond correlations are seen, allowing the identification of adjacent atoms.

By adjusting the mixing time and allowing the spins to diffuse further, correlations between more distant nuclei may be observed and nuclei of neighboring residues being identified. Exact mixing times will vary between sample types but good starting points are ~ 20 ms for one-bond correlations, ~ 100 ms for complete side chains up to 500 ms for inter-residue correlations. Variation of mixing times in broadband dipolar recoupling experiments allows distance restraints to be acquired to assist structure calculation.

The pulse sequence for a PDSD experiment is implemented as follows: the proton magnetization is excited with a 90° pulse and magnetization transferred to the carbon-13 by means of ramped cross-polarization. The magnetization on the ^{13}C is then allowed to evolve during t_1 under a high-power ^1H decoupling scheme, such as SPINAL-64 proton decoupling [34]. A 90° pulse on the ^{13}C channel subsequently stores the magnetization along the z -axis and the proton decoupling is turned off during the mixing period, the length of which is varied in order to acquire either short or longer range ^{13}C – ^{13}C correlations. At the end of the mixing period a second 90° pulse restores the magnetization back to the transverse plane and the decay of the magnetization (known as the free induction decay or FID) is observed during the period t_2 , again under high-power proton decoupling. A schematic of the pulse sequence can be seen in Fig. 3.

4.3 Two-Dimensional Heteronuclear Correlation Experiments

HXYCP experiments are used to correlate different low γ nuclei within a protein. They are analogous to liquid-state triple-resonance experiments, although ^{13}C rather than ^1H spectra are detected. During NCA/NCO experiments, magnetization is transferred from ^1H to ^{15}N by a ramped CP pulse, and the chemical shift is

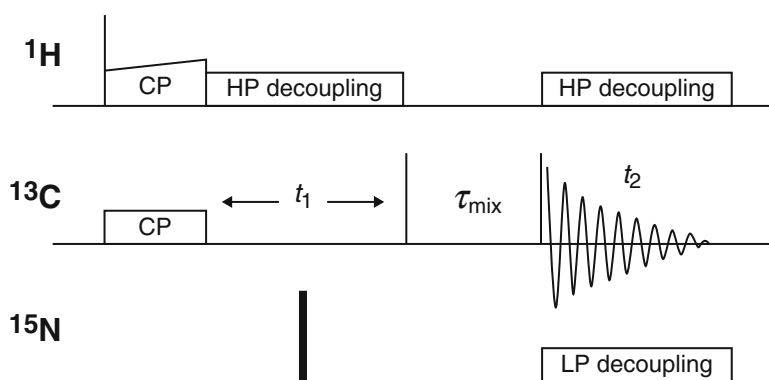


Fig. 3 Pulse sequence for cross-polarized proton-driven spin diffusion NMR experiment. The *top line* represents pulses applied on the proton frequency and the *bottom line* pulses applied on the ^{13}C frequency. The *narrow black rectangles* represent 90° applied pulses. The first evolution period is t_1 . *Ramped boxes* containing "CP" are contact pulses for cross-polarization. t_{mix} is the mixing time. The FID is represented by t_2 . Decoupling pulses are represented by *boxes* marked decoupling and are denoted HP (high power) and LP (low power)

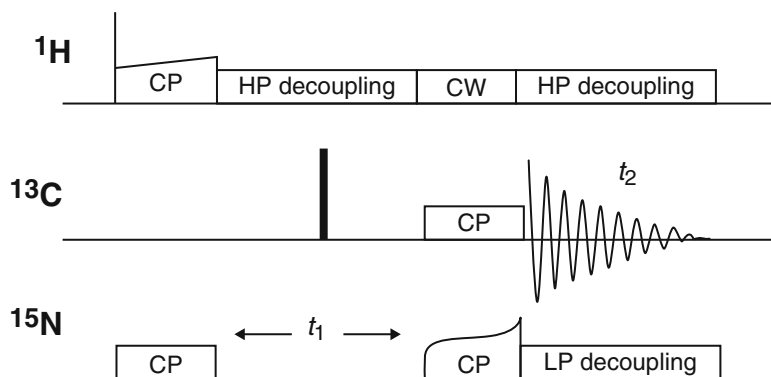


Fig. 4 Pulse sequence for a specific HXYCP NMR experiment. The *top line* represents pulses applied on the proton frequency, the *bottom line* ^{15}N , and the *middle line* pulses applied on either the $^{13}\text{C}\alpha$ or $^{13}\text{C}'$ frequency depending on the experiment. The *narrow black rectangles* represent 90° applied pulses. *Ramped boxes* containing “CP” are contact pulses for cross-polarization. t_1 is the first evolution period. Decoupling pulses are represented by *boxes* marked decoupling. Boxes marked “Specific CP” represent low-power contact pulses from specific cross-polarization. The FID is represented by t_2

evolved on the ^{15}N nuclei under high-power proton decoupling during the period t_1 . Instead of magnetization being transferred from ^{15}N to all connecting ^{13}C nuclei as in the classic HXY-CP experiment [35] it is selectively transferred to either the $^{13}\text{C}\alpha$ or $^{13}\text{C}'$ using specific cross-polarization. This is achieved through irradiation of selected areas of the ^{13}C spectrum with a low-power (30 kHz) radio frequency contact pulse while other areas dipphase. Thus directed transfer of magnetization and improved signal-to-noise ratio is allowed as magnetization is not split between both of the sites [36]. During specific cross-polarization, proton decoupling can be applied using only continuous-wave decoupling to reduce any cross-polarization to the protons through the presence of phase transients, or alternatively it may be omitted completely. An FID is then acquired from the ^{13}C spins during the period t_2 under the usual preferred proton decoupling scheme. A schematic of an HXY-CP two-dimensional experiment can be seen in Fig. 4. These experiments are used in conjunction with the PDSD experiment to sequentially assign proteins.

4.4 NCACX and NCOCX Experiments

The NCACX and NCOCX experiments are commonly used for the identification of spin systems and the sequential assignment of resonances. They are “three-dimensional” experiments which decrease spectral crowding and therefore are particularly effective for large proteins whose two-dimensional spectra are too crowded to yield a substantial assignment. The NCACX experiment correlates the amide nitrogen with the aliphatic carbons, while the

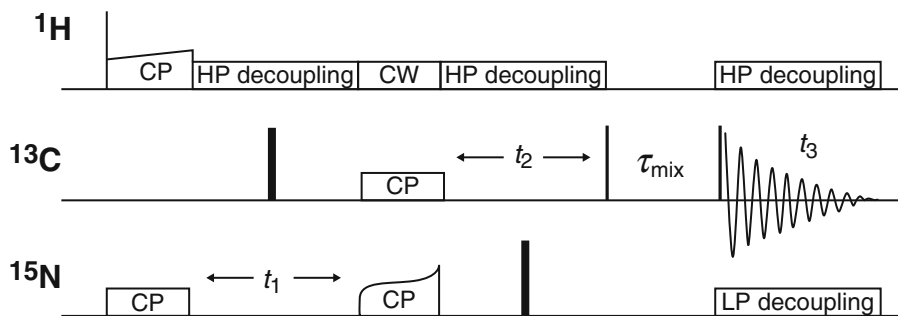


Fig. 5 Pulse sequence for an NCACX/NCOCX 3D NMR experiment. The *top line* represents pulses applied on the proton frequency, the *bottom line* ^{15}N , and the *middle line* pulses applied on either the $^{13}\text{C}\alpha$ or $^{13}\text{C}'$ frequency depending on the experiment. The *narrow black rectangles* represent 90° applied pulses. *Ramped boxes* containing “CP” are contact pulses for cross-polarization. t_1 is the first evolution period. Decoupling pulses are represented by *boxes* marked decoupling. *Boxes* marked “Specific CP” represent low-power contact pulses from specific cross-polarization. t_2 is the second evolution period. t_{mix} is mixing time. The FID is represented by t_3

NCOCX experiment correlates the amide nitrogen with intra-residue carbonyl carbon and aliphatic carbons of the previous residue, allowing for sequential assignment and disambiguation of resonances. The first part of these experiments proceeds exactly as the HXY CP experiment, with magnetization transferred to the ^{15}N spins before being allowed to evolve under the ^{15}N chemical shift with high-power proton decoupling before being transferred to either $^{13}\text{C}\alpha$ or $^{13}\text{C}'$ spins via specific CP. At this point, magnetization is allowed to evolve on the $^{13}\text{C}\alpha/^{13}\text{C}'$ under proton decoupling. After evolution, a PDSD step is used to propagate magnetization to neighboring ^{13}C nuclei and an FID is then acquired under proton decoupling in the period t_3 . Figure 5 shows the pulse sequence for the NCACX/NCOCX experiment. A full and in-depth account for the experiment was published by Jutta Pauli in 2001 [37].

To summarize, the first experiments that should be acquired for structure determination are a series of broadband recoupling measurements with different mixing times. We recommend the PDSD for its simplicity or the DARR experiment for its transfer efficiency at longer mixing times. Then, NCA and NCO experiments should be used for intra- and inter-residue correlation of nitrogens to carbons so that sequential assignments can be made. Ideally, if signal to noise allows, NCACX and NCOCX experiments should be acquired and this is certainly essential for larger membrane proteins that produce more crowded spectra.

4.5 Generating Angle Restraints from ssNMR Chemical Data

The chemical shift of a nucleus (measured in ppm) is highly sensitive to the asymmetrical environment of the electron cloud that surrounds it, which in turn is influenced by the neighboring atoms and their geometry [38]. Once peak assignment is complete, the

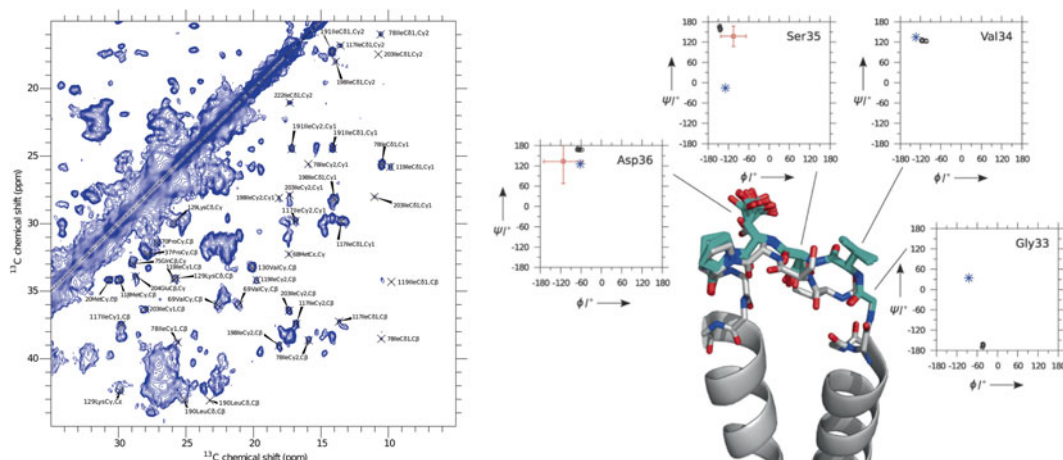


Fig. 6 Analysis of bacteriorhodopsin by solid-state NMR. *Left*: DARR spectrum of U- ^{13}C - ^{15}N bacteriorhodopsin in the purple membrane acquired at 800 MHz with a MAS speed of 10.7 kHz and at -10°C . Site-specific assignments are shown for resolved residues. *Right*: The structure of the AB loop of bacteriorhodopsin: The 1QHJ crystal structure is shown in grey, and the ten lowest energy structures calculated using XPLOR based on the dihedral angles predicted by TALOS+ are shown in cyan. Side-chain conformations and Ramachandran plots are shown for Gly33–Pro37: 1QHJ crystal structure (blue stars), TALOS+ prediction (red squares), ensemble structure calculated by XPLOR-NIH (black circles). Figure adapted from Higman et al. [42]

dihedral φ and ψ angles of individual amino acids may be determined from the secondary chemical shifts ($\Delta\delta$) of individual resonances, where $\Delta\delta = \Delta\delta_{\text{observed}} - \Delta\delta_{\text{random coil}}$. $\text{C}\alpha$ atoms in alpha helical protein regions generally have a $\Delta\delta > 0$, whereas $\text{C}\alpha$ atoms from identical residue types in β stranded regions will have a $\Delta\delta < 0$. The pattern of secondary chemical shifts is reversed for the $\text{C}\beta$ atom of an amino acid, so in α helical and β stranded regions, the $\Delta\delta$ for $\text{C}\beta$ is negative and positive, respectively [39, 40]. The dihedral angles may be determined by comparing the measured secondary chemical shifts to those in published databases. Several such databases exist, although TALOS+ is most commonly used for membrane proteins [41]. Figure 6 shows how chemical shifts measured in a DARR spectrum were used in combination with the TALOS+ database to generate the structure of an interhelical loop of the 7TM photoreceptor bacteriorhodopsin from *Halobacterium salinarum* [42].

5 Oriented Sample (OS) Methods

MAS NMR experiments eliminate the anisotropic line broadening that is not averaged by rapid sample tumbling, and so the peak positions observed in the spectra are equivalent to those that would be observed in solution. An alternative approach, which allows the

angle-dependent chemical shift and dipolar couple to be retained (and not averaged by MAS), requires membrane samples to be mechanically or magnetically aligned. In this way, the labeled proteins are present in only one orientation throughout the sample relative to the direction of the magnetic field. Mechanical orientation is typically achieved by layering of the membrane samples on stacked glass plates [27], although orientation of bicelles and nanodiscs using conjugated paramagnetic ions (including lanthanides) is increasingly common [5, 23].

A common application of OS NMR is the determination of transmembrane and membrane-surface-bound alpha helix orientation. The backbone N–H bonds, which play a key role in secondary structure stabilization, lie approximately parallel to the helical axis. The ^{15}N chemical shift anisotropy tensors have a ($\Delta\sigma$) magnitude of ~ 160 ppm [43]. When a helix is aligned with its long axis parallel to the membrane, the backbone ^{15}N resonances appear around 200 ppm in a ^{15}N spectrum, whereas resonances below 100 ppm are recorded when the same axis is perpendicular to the magnetic field [44]. More complex 2D spectra (including PISEMA and HETCOR experiments), which exploit the anisotropy of the ^1H – ^{15}N dipolar couplings, allow the assignment of individual resonances and allow local orientations to be determined; PISEMA provides better resolution of transmembrane helices; and HETCOR is more suitable for studying intra- and extracellular loop regions of the polypeptide chain [45, 46]. 2D oriented spectra are routinely used to determine orientations of short peptides or longer polytopic proteins including bacteriorhodopsin [16]. A low E-field square-coil probe for oriented samples is shown in Fig. 2 [47].

Deuterated groups undergoing rapid anisotropic motion give rise to two peaks in a ^2H NMR spectrum, as a result of the quadrupole moment of the spin 1 deuterium nucleus [48]. The quadrupole splitting ($\Delta\nu_Q$) between the two peaks may be analyzed to determine the orientation with respect to the membrane of C–CD₃ bonds in site-specifically labeled proteins and peptides [49]. Alanine residues, deuterated at C β , are placed at strategic points along an alpha helical section of peptide which is reconstituted into mechanically aligned bilayers and the effect on sequence changes on the membrane orientation of model Trp-Ala-Leu-Pro peptides has been studied by this method [50]. NMR studies of CD₃-labeled retinal have allowed differences in chromophore in both bovine rhodopsin and bacteriorhodopsin to be identified [51, 52]. The orientation of C–CD₃ groups in retinal in both bacteriorhodopsin photocycle intermediates has been determined by ^2H magic angle oriented sample spinning (MAOSS) NMR, which exploits the orientational dependence of the position of ^2H spinning side bands [53].

6 Fast MAS Experiments

^{13}C -detected experiments at fast MAS such as ^{15}N – ^{13}C correlations or ^{13}C – ^{13}C correlations are, in principle, similar to the equivalent experiments at slow or moderate MAS frequencies. However, the more efficient reduction in heteronuclear dipolar couplings has several consequences. Firstly, efficient low-power ^1H decoupling schemes are accessible, and provided that corresponding low-power cross-polarization (CP) transfer conditions are found, sample heating due to the RF pulses employed during signal acquisition is markedly reduced. This makes the interscan delay a function only of the bulk ^1H T_1 which can even be decreased by the use of paramagnetic agents. Secondly, PDSF transfers between ^{13}C nuclei become inefficient at fast MAS speeds and so homonuclear recoupling schemes such as DREAM are used instead. In addition, the larger separation between Hartmann-Hahn conditions for CP magnetization transfer facilitates specific transfers to carbonyl or aliphatic carbons. The longer coherence lifetimes also make scalar coupling-based transfers a viable alternative, providing much more control over magnetization evolution [54].

In addition to allowing the faster acquisition of ^{13}C -detected experiments, fast MAS frequencies allow for spectral acquisition using the more sensitive ^1H nuclei. The advantages of ^1H detection are not just increased sensitivity but also the ability to measure and correlate ^1H chemical shifts which reduces much of the ambiguity in resonance assignment. This can be done using the well-established protocols developed for solution NMR where ^1H detection is the norm. The current suite of ^1H -detected solid-state NMR pulse programs seeks to emulate the “standard” experiments used in solution, with equivalents to the HNCA, HN(CO)CA, HNCO, HN(CA)CO, HN(CA)CB, and HN(COCA)CB experiments frequently employed for backbone resonance assignment. Once the signals are assigned, experiments that recouple ^1H spins allow information on internuclear distances to be acquired, which can be then used for structure calculation.

Although the use of low-power pulse sequences renders RF heating negligible, frictional heating due to the fast rotation of the sample rotor is significant, and increases nonlinearly with MAS frequency [55]. A 1.3 mm rotor spinning at 60 kHz typically causes the temperature of the sample to be anywhere from 30 to 70 °C warmer than that reported by the temperature control unit on the spectrometer. The difference between the reported and actual temperatures is dependent on the position of the thermocouple within the probe. Therefore, it is recommended to calibrate the temperature on each probe using a rotor containing a material which exhibits chemical shifts or relaxation rates which are temperature dependent, such as KBr [56], $\text{Pb}(\text{NO}_3)_2$ [57], $\text{Sm}_2\text{Sn}_2\text{O}_7$ [58], or even water [55].

7 Conclusion

Solid-state NMR is a rapidly developing, versatile technique, capable of producing high-resolution structural information for functionally active integral membrane proteins in a native or near-native environment. It enhances our understanding of functional mechanisms by providing insights into protein dynamics and membrane interactions. Its strength as a biophysical methodology lies in the ability of the data produced to be combined with other techniques (both experimental and computational) and it therefore has a pivotal role to play in the study of this fascinating and challenging group of proteins.

Acknowledgements

The authors acknowledge funding from the Medical Research Council (UK), the Biotechnology and Biological Sciences Research Council (UK), the European Metrology Research Programme, and the National Physical Laboratory (London, UK).

References

1. Wallin E, von Heijne G (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci* 7:1029–1038
2. Boyd D, Schierle C, Beckwith J (1998) How many membrane proteins are there? *Protein Sci* 7:201–205
3. Russell RB, Eggleston DS (2000) New roles for structure in biology and drug discovery. *Nat Struct Biol* 7(Suppl):928–930
4. Watts A (2005) Solid-state NMR in drug design and discovery for membrane-embedded targets. *Nat Rev Drug Discov* 4:555–568
5. Judge PJ, Watts A (2011) Recent contributions from solid-state NMR to the understanding of membrane protein structure and function. *Curr Opin Chem Biol* 15:690–695
6. Kamiyama M, Vosegaard T, Mason AJS (2005) Structural and orientational constraints of bacteriorhodopsin in purple membranes determined by oriented-sample solid-state NMR spectroscopy. *J Struct Biol* 149:7–16
7. Zhou DH, Shah G, Mullen C et al (2009) Proton-detected solid-state NMR spectroscopy of natural-abundance peptide and protein pharmaceuticals. *Angew Chem* 48: 1253–1256
8. Chevelkov V, Rehbein K, Diehl A, Reif B (2006) Ultrahigh resolution in proton solid-state NMR spectroscopy at high levels of deuteration. *Angew Chem* 45:3878–3881
9. Linser R, Fink U, Reif B (2008) Proton-detected scalar coupling based assignment strategies in MAS solid-state NMR spectroscopy applied to perdeuterated proteins. *J Magn Reson* 193:89–93
10. Akbey U, Lange S, Trent Franks W (2010) Optimum levels of exchangeable protons in perdeuterated proteins for proton detection in MAS solid-state NMR spectroscopy. *J Biomol NMR* 46:67–73
11. Hong M, Jakes K (1999) Selective and extensive ^{13}C labeling of a membrane protein for solid-state NMR investigations. *J Biomol NMR* 14:71–74
12. Castellani F, van Rossum B, Diehl A et al (2002) Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy. *Nature* 420:98–102
13. Becker J, Ferguson N, Flinders J et al (2008) A sequential assignment procedure for proteins that have intermediate line widths in MAS NMR spectra: Amyloid fibrils of human CA150. *ChemBiochem* 9:1946–1952

14. Cross TA, Sharma M, Yi M, Zhou HX (2011) Influence of solubilizing environments on membrane protein structures. *Trends Biochem Sci* 36:117–125
15. Murray DT, Das N, Cross TA (2013) Solid State NMR strategy for characterizing native membrane protein structures. *Acc Chem Res* 46:2172–2181
16. Judge PJ, Taylor GF, Vermeer LS, Watts A (2014) Structural insights from solid-state NMR into the function of the bacteriorhodopsin photoreceptor protein. In: Separovic F, Naito A (eds) *Advances in biological solid-state NMR: proteins and membrane-active peptides*. The Royal Society of Chemistry, Cambridge, UK
17. Warschawski DE, Arnold AA, Beaugrand M et al (2011) Choosing membrane mimetics for NMR structural studies of transmembrane proteins. *Biochim Biophys Acta* 1808:1957–1974
18. Rigaud JL, Levy D (2003) Reconstitution of membrane proteins into liposomes. *Meth Enzymol* 372:65–86
19. Das N, Murray DT, Cross TA (2013) Lipid bilayer preparations of membrane proteins for oriented and magic-angle spinning solid-state NMR samples. *Nat Protoc* 8:2256–2270
20. Abdine A, Park KH, Warschawski DE (2012) Cell-free membrane protein expression for solid-state NMR. *Meth Mol Biol* 831:85–109
21. Abdine A, Verhoeven MA, Warschawski DE (2011) Cell-free expression and labeling strategies for a new decade in solid-state NMR. *New Biotechnol* 28:272–276
22. Prosser RS, Evanics F, Kitevski JL, Al-Abdul-Wahid MS (2006) Current applications of bicelles in NMR studies of membrane-associated amphiphiles and proteins. *Biochemistry* 45:8453–8465
23. Diller A, Loudet C, Aussenac F et al (2009) Bicelles: a natural “molecular goniometer” for structural, dynamical and topological studies of molecules in membranes. *Biochimie* 91:744–751
24. Cho HS, Dominick JL, Spence MM (2010) Lipid domains in bicelles containing unsaturated lipids and cholesterol. *J Phys Chem B* 114:9238–9245
25. Park SH, Opella SJ (2010) Triton X-100 as the “short chain lipid” improves the magnetic alignment and stability of membrane proteins in phosphatidylcholine bilayers for oriented sample (OS) solid-state NMR Spectroscopy. *J Am Chem Soc* 132:12552–12553
26. Marcotte I, Belanger A, Auger M (2006) The orientation effect of gramicidin A on bicelles and Eu³⁺-doped bicelles as studied by solid-state NMR and FT-IR spectroscopy. *Chem Phys Lipids* 139:137–149
27. Grobner G, Taylor A, Williamson PT et al (1997) Macroscopic orientation of natural and model membranes for structural studies. *Anal Biochem* 254:132–138
28. Varga K, Watts A (2007) Introduction to solid state NMR and its application to membrane protein-ligand binding studies. In: Pebay-Peyroula E (ed) *Biophysical analysis of membrane proteins. Investigating structure and function*. Wiley-VCH, Weinheim, pp 55–87
29. Bertini I, Engelke F, Luchinat C et al (2012) NMR properties of sedimented solutes. *Phys Chem Chem Phys* 14:439–447
30. Bockmann A, Gardienet C, Verel R et al (2009) Characterization of different water pools in solid-state NMR protein samples. *J Biomol NMR* 45:319–327
31. Watts A, Straus SK, Grage S et al (2004) Membrane protein structure determination using solid state NMR. In: Downing K (ed) *Methods in molecular biology – techniques in protein NMR*, vol 278. Humana Press, Totowa, NJ, pp 403–474
32. Takegoshi K, Nakamura S, Terao T (2001) C-13-H-1 dipolar-assisted rotational resonance in magic-angle spinning NMR. *Chem Phys Lett* 344:631–637
33. Bayro MJ, Ramachandran R, Caporini MA et al (2008) Radio frequency-driven recoupling at high magic-angle spinning frequencies: homonuclear recoupling sans heteronuclear decoupling. *J Chem Phys* 128:052321
34. Fung BM, Khitrin AK, Ermolaev K (2000) An improved broadband decoupling sequence for liquid crystals and solids. *J Magn Reson* 142:97–101
35. Baldus M, Geurts DG, Hediger S, Meier BH (1996) Efficient 15N–13C polarization transfer by adiabatic-passage Hartmann–Hahn cross polarization. *J Magn Reson A* 118:140–144
36. Baldus M, Petkova A, Herzfeld J, Griffin R (1998) Cross polarization in the tilted frame: assignment and spectral simplification in heteronuclear spin systems. *Mol Phys* 95:1197–1207
37. Pauli J, Baldus M, van Rossum B, de Groot H et al (2001) Backbone and side-chain 13C and 15N signal assignments of the alpha-spectrin SH3 domain by magic angle spinning solid-state NMR at 17.6 Tesla. *ChemBiochem* 2:272–281
38. Higman VA (2013) Proteins in solution and at interfaces. In: Pineiro A, Ruso J (eds) *Methods and applications in biotechnology and materials science*. Wiley-Blackwell, New York, pp 23–48

39. Wishart DS, Sykes BD (1994) The ^{13}C chemical-shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. *J Biomol NMR* 4:171–180
40. Wishart DS, Sykes BD, Richards FM (1992) The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry* 31:1647–1651
41. Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS plus: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J Biomol NMR* 44: 213–223
42. Higman VA, Varga K, Aslimovska L et al (2011) The conformation of bacteriorhodopsin loops in purple membranes resolved by solid-state MAS NMR spectroscopy. *Angew Chem* 50:8432–8435
43. Yao L, Grishaev A, Cornilescu G, Bax A (2010) Site-specific backbone amide (^{15}N) chemical shift anisotropy tensors in a small protein from liquid crystal and cross-correlated relaxation measurements. *J Am Chem Soc* 132: 4295–4309
44. Bechinger B, Sizun C (2003) Alignment and structural analysis of membrane polypeptides by ^{15}N and ^{31}P solid-state NMR spectroscopy. *Concepts Magn Reson* 18A:130–145
45. Vosegaard T, Nielsen NC (2002) Towards high-resolution solid-state NMR on large uniformly ^{15}N - and $[^{13}\text{C},^{15}\text{N}]$ -labeled membrane proteins in oriented lipid bilayers. *J Biomol NMR* 22:225–247
46. Marassi FM, Ma C, Gesell JJ, Opella SJ (2000) Three-dimensional solid-state NMR spectroscopy is essential for resolution of resonances from in-plane residues in uniformly (^{15}N) N-labeled helical membrane proteins in oriented lipid bilayers. *J Magn Reson* 144: 156–161
47. Gor'kov PL, Witter R, Chekmenev EY et al (2007) Low-E probe for ^{19}F – ^1H NMR of dilute biological solids. *J Magn Reson* 189:182–189
48. Seelig J (1977) Deuterium magnetic resonance: theory and application to lipid membranes. *Q Rev Biophys* 10:353–418
49. Ulrich AS, Wallat I, Heyn MP, Watts A (1995) Re-orientation of retinal in the M-photointermediate of bacteriorhodopsin. *Nat Struct Biol* 12:190–192
50. Vostrikov VV, Daily AE, Greathouse DV, Koeppe RE II (2010) Charged or aromatic anchor residue dependence of transmembrane peptide tilt. *J Biol Chem* 285:31723–31730
51. Salgado GF, Struts AV, Tanaka K et al (2004) Deuterium NMR structure of retinal in the ground state of rhodopsin. *Biochemistry* 43:12819–12828
52. Ulrich AS, Watts A, Wallat I, Heyn MP (1994) Distorted structure of the retinal chromophore in bacteriorhodopsin resolved by ^2H -NMR. *Biochemistry* 33:5370–5375
53. Glaubitz C, Burnett IJ, Grobner G et al (1999) Deuterium MAS NMR spectroscopy on oriented membrane proteins: applications to photointermediates of bacteriorhodopsin. *J Am Chem Soc* 121:5787–5794
54. Vijayan V, Demers JP, Biernat J et al (2009) Low-power solid-state NMR experiments for resonance assignment under fast magic-angle spinning. *Chemphyschem* 10:2205–2208
55. Dvinskikh SV, Castro V, Sandstrom D (2004) Heating caused by radiofrequency irradiation and sample rotation in ^{13}C magic angle spinning NMR studies of lipid membranes. *Magn Reson Chem* 42:875–881
56. Thurber KR, Tycko R (2009) Measurement of sample temperatures under magic-angle spinning from the chemical shift and spin-lattice relaxation rate of ^{79}Br in KBr powder. *J Magn Reson* 196:84–87
57. Bielecki A, Burum DP (1995) Temperature dependence of ^{207}Pb MAS spectra of solid lead nitrate. An accurate, sensitive thermoMETER for variable-temperature MAS. *J Magn Reson A* 116:215–220
58. Langer B, Schnell II, Spiess HW, Grimmer AR (1999) Temperature calibration under ultrafast MAS conditions. *J Magn Reson* 138:182–186

Chapter 18

Native Mass Spectrometry: Towards High-Throughput Structural Proteomics

Frances D.L. Kondrat, Weston B. Struwe, and Justin L.P. Benesch

Abstract

Native mass spectrometry (MS) has become a sensitive method for structural proteomics, allowing practitioners to gain insight into protein self-assembly, including stoichiometry and three-dimensional architecture, as well as complementary thermodynamic and kinetic aspects. Although MS is typically performed in vacuum, a body of literature has described how native solution-state structure is largely retained on the timescale of the experiment. Native MS offers the benefit that it requires substantially smaller quantities of a sample than traditional structural techniques such as NMR and X-ray crystallography, and is therefore well suited to high-throughput studies. Here we first describe the native MS approach and outline the structural proteomic data that it can deliver. We then provide practical details of experiments to examine the structural and dynamic properties of protein assemblies, highlighting potential pitfalls as well as principles of best practice.

Key words Mass spectrometry, Gas-phase protein structure, Ion mobility spectrometry, Collision cross section

1 Introduction

High-throughput DNA sequencing has led to an explosion in the amount of genomic information available, providing insight into the differences between organisms, individuals, and cell types, as well as identifying candidate mutations for many human diseases [1]. Traditional structural biology approaches have found it difficult to keep pace, and consequently translating the abundance of genomic data into knowledge of the structure and dynamics of the gene products represents a significant bottleneck in modern bioscience. Therefore, robust technologies are needed that can act to close this gap, providing structural information on all levels of protein organization, and to do so in a high-throughput manner.

Mass spectrometry (MS) has in the last 20 years transformed the field of proteomics: it is now possible to rapidly identify proteins, characterize their posttranslational modifications, and even

quantify the abundance of essentially complete proteomes from mammalian cell lines [2]. In parallel to this, MS technologies and methodologies have been developed that enable the study of protein structure [3, 4]. Perhaps most surprisingly, under experimental and instrumental conditions chosen to preserve non-covalent interactions, MS has emerged as an effective means for interrogating the quaternary architecture of proteins [5–9].

Native MS can be used to study some of the most challenging of protein properties, with intrinsic disorder [10], polydispersity [3], and association with membranes [11] all proving to be surmountable. This broad applicability, combined with the use of relatively small quantities of sample and the rapidity of analysis, means that native MS is well placed for high-throughput structural biology studies [12].

1.1 Native Mass Spectrometry

The term “native MS” was coined by analogy to native polyacrylamide gel electrophoresis [13], and its success relies on both the development of “soft” ionization techniques and mass spectrometers capable of transmitting large biomolecular ions [14]. The implementation of this approach can be dated back two decades [15, 16], to studies that reported the preservation of intact protein assemblies or protein-ligand complexes in the gas phase [17]. This early work demonstrated that the unparalleled mass accuracy afforded by MS enabled the unambiguous determination of protein stoichiometry, simply through comparison to the mass of the constituents. By the turn of the century, complexes on the MDa scale could be accurately mass-measured, and today the “record” stands at an astonishing 18 MDa [18].

Aside from having the potential to be very large, protein assemblies can also be extremely intricate machines [19]. Accordingly, to address this complexity, methodologies have been developed that enable manipulation of the protein assemblies either prior to their introduction into the mass spectrometer, or within the instrument, to tease out structural details beyond simply stoichiometry. The utility of these experiments rests on the observation that, on the rapid timescale of native MS measurements, protein structure in the gas phase strongly reflects that in solution [20, 21]. As a result, native MS is an exciting method for interrogating protein assemblies: as a screening approach prior to downstream analyses, as a source of data to be combined with that from other approaches, or as a means for determining protein structure in its own right [3–9].

1.2 Generating and Transmitting Ions of Intact Protein Assemblies

Native MS experiments rely on electrospray ionization (ESI), a process whereby a solution of proteins is held in a conductive capillary and aerosolized at atmospheric pressure into a stream of charged droplets by the application of a potential relative to the inlet of the mass spectrometer [22]. These droplets can be either

negatively or positively charged, depending on the ESI polarity used. Solvent evaporation from the droplets leads to their shrinking until the Rayleigh limit is reached, and fission events occur. These processes continue until all solvent is depleted, and only the native protein ions, and other involatile components of the solution, remain [23]. ESI therefore essentially amounts to a dehydration process, and represents a gentle means for generating protein ions. A miniaturized form, nanoESI [24], is particularly attractive for native MS, as the small capillary diameter results in low flow rates (leading to typically picomole sample quantities), as well as small initial droplet sizes that allow the use of aqueous buffers at relatively mild interface conditions [14]. Nevertheless, in practice the introduction of sample into the instrument remains arguably the most important step in the native MS experiment and hinges on employing buffer conditions compatible both with protein integrity in solution and stable ESI.

Once generated, the ions must be focussed into a beam to travel through the various stages of the mass spectrometer until they impact on the detector. This relies on optimizing pressures and potentials, particularly in the early vacuum stages, to ensure successful transmission of protein complexes [14]. These parameters are instrument dependent, and can also be tuned to improve the quality of the mass spectra, or to cause intentional activation of the protein assemblies. These possibilities however demonstrate that parameters can also be easily overlooked, resulting in inadvertent dissociation of the complexes. Therefore considerable care needs to be taken to ensure that non-covalent interactions are preserved and measurements of native structures are made.

1.3 Ion Activation in the Gas Phase

The ability to intentionally induce the gas-phase activation of ions within the mass spectrometer is invaluable in the study of protein assemblies [25]. In the study of homomeric assemblies, mass measurement of the intact ions is often sufficient to unambiguously determine stoichiometry. However, the presence of multiple components within many protein assemblies often means that their stoichiometry cannot be deduced from intact mass alone. It is therefore useful to disrupt the non-covalent bonds holding these complexes together, allowing the mass measurement of their constituents. A loose analogy may be drawn between this approach, and the more familiar use of MS in peptide sequencing, in which gas-phase dissociation is used to break apart the covalent bonds along the peptide backbone, yielding fragments from which primary sequence data is obtained.

A number of different gas-phase activation techniques exist, and can be applied in a wide range of different mass spectrometer configurations [14]. The most common technique applied to the study of protein complexes is collision-induced dissociation (CID), in which activation is promoted by their acceleration into multiple collisions with inert gas molecules [26]. These collisions act to

increase the internal energy of the complex ion, ultimately leading to the expulsion of one or more subunits, in a sequential manner [27]. The released subunit is typically unfolded and, by virtue of this increased surface area, can accommodate a large number of charges [28]. This leads to characteristic “asymmetric dissociation,” with the resultant spectra featuring two important regions: charge states corresponding to highly charged subunits at low m/z and the residual “stripped” complexes carrying relatively few charges and appearing at high m/z [25].

1.3.1 Exploiting CID to Measure Stoichiometry

Implementing CID can be a very profitable strategy when attempting to determine protein stoichiometry, as different combinations of the constituent subunits can often be consistent with the mass of the complex [29]. Firstly, accurately measuring the mass of subunits released from the complex may reveal discrepancies between the measured data and primary sequence information, due to posttranslational modification. Secondly, in cases where assignment is difficult for the complex (with n subunits), measuring the mass of the stripped complex (with $n-1$ subunits) can prove a useful alternative. In this way, by accounting for the expelled subunit(s), the mass of the parent may be back-calculated [30]. This approach is particularly useful in the case of polydisperse samples, where there is a lot of peak overlap in the mass spectra. Because CID results in the removal of subunits with a disproportionate amount of charge, the peaks corresponding to stripped species are separated in the mass spectrum to a greater extent with respect to the intact complex [27].

In addition to obtaining information on composition, the CID pathway can provide some insights into the three-dimensional organization of protein complexes [31]. Unfortunately, however, certain protein complexes do not break apart readily upon collisional activation, or, in so doing, do not allow all subunits within the heteromer to be identified. Generally small and exposed subunits are expelled first and so are most frequently observed, whereas larger and buried subunits are often missing from the spectra [25]. In these instances solution-phase perturbation can often be fruitfully combined with CID to gather sufficient data on the complex [29], or other activation approaches may be employed [32].

1.4 Measuring Size as well as Mass by means of Ion Mobility Mass Spectrometry

The introduction of commercial mass spectrometers incorporating ion mobility (IM) separation is a relatively recent development, and has provided an additional dimension of separation that is capable of providing insight into the three-dimensional structure of proteins [33]. Different means to achieve IM separation have been developed, but in essence all are based on exploiting the charge and physical size of the analyte [34]. In conjunction with MS, the orthogonal dimension of separation imparted by IM causes a dramatic increase in peak capacity, leading to the ability to

resolve molecules that overlap in m/z [35]. Furthermore, it is possible to guide the modeling of protein complexes with IM-MS data in circumstances where conventional structural biology approaches have proven intractable [36, 37].

1.4.1 Determining Collisional Cross Sections

The most common forms of IM separation employed in native MS measure the time taken for the ion to traverse a neutral gas under the influence of a weak electric field [38]. This drift time is related to the rotationally averaged collisional cross section (CCS, Ω), which can be used to report on the shape of the ion. The application of IM-MS to structural studies of proteins has seen a rapid growth due to the availability of commercial hybrid Q-IM-ToF instruments built on “travelling-wave” technology [39], in which ions are propelled through the IM gas by waves of direct current. In this experiment, calibration using protein standards of known CCS is necessary in order to obtain accurate values. However, since ultimately CCS is an absolute quantity, depending only on the analyte ion and target gas, it can be measured on a range of instruments, directly compared, and employed as a direct restraint for structural modeling.

1.4.2 Monitoring Gas-Phase Unfolding

The disruption of non-covalent interactions between protein complex subunits in the mass spectrometer can provide a certain amount of information regarding the strength of ligand binding, or to identify changes in variant forms. This can be done by determining the energies that affect CID [40], or, in a method that provides richer information, examining the unfolding that precedes dissociation [41, 42]. To perform this latter experiment, progressive gas-phase unfolding is effected by collisional activation, and monitored by means of IM, frequently revealing intermediates with different CCS along the unfolding trajectory. The energy at which transitions between these native and intermediate states are observed represents a characteristic “fingerprint” of the protein under study [43].

1.5 Monitoring Dynamics of Protein Assemblies

Proteins are naturally dynamic entities, converting between different conformations both at equilibrium, and upon stimulation. In addition to the canonical fluctuations in secondary and tertiary structure, protein complexes also have the ability to gain, lose, or reorganize subunits. These quaternary dynamics underpin the self-assembly and disassembly processes, yet their functional implications have not been studied in as much detail as warranted, due to the experimental challenges involved. Native MS has in the last few years emerged as the biophysical method of choice to quantify quaternary dynamics, due largely to the ability of MS to observe each individual species within a sample, in real time, where other biophysical techniques only provide a report on the solution average.

As MS has the ability to observe all species concurrently, experiments can be conducted in which multiple components are mixed and complex formation followed [44]. As these experiments can be conducted in real time they may report on stable intermediates formed in the assembly process. It is also possible to study the dynamics of homomeric complexes at equilibrium by equilibrating subunits with labeled counterparts [45, 46]. Such a subunit exchange experiment can not only reveal the kinetics, but also provide information regarding the exchanging species [47]. Furthermore, extracting and comparing the rates of exchange between wild-type and mutant constructs allow specific residues or regions involved in the dynamics of protein complexes to be highlighted [46, 48].

2 Materials

2.1 *Mass Spectrometer*

1. Analysis of protein assemblies, intact in the gas phase, by means of native MS in general requires specialized instrumentation. It is now possible to buy certain models of mass spectrometer directly from the manufacturer with the capability of performing such experiments, specifically Synapt (Waters) and Exactive Plus EMR (ThermoFisher Scientific) instruments. An alternative is provided by companies that recondition mass spectrometers, modifying them specifically for native MS applications (MS Vision, MS Service Solutions).
2. Time-of-flight (ToF) instruments are the simplest instruments for native MS, and in general provide good-quality mass spectra at low cost [49, 50]. Collisional activation of all the ions can be effected in the source region of the instrument, but is not mass selective.
3. Orbitrap analyzers offer an alternative to ToF, and benefit from intrinsic higher mass resolution and a reduction in chemical noise [51]. The Exactive Plus EMR (ThermoFisher Scientific) allows collisional activation both in source and in a dedicated collision cell (without prior mass selection) that combine to provide efficient desolvation [52].
4. Quadrupole-ToF (Q-ToF) instruments are currently the most popular choice for native tandem-MS, combining the benefits of the high mass range attainable with ToF analyzers, with the ability for mass-selective CID [53–55]. The Synapt instruments (Waters) are a variant of the Q-ToF geometry, incorporating IM separation, thereby providing size information to complement the mass measurement [56].
5. Fourier transform ion cyclotron resonance (FTICR) instruments provide the highest resolution of all, even allowing for

isotopic resolution of protein complexes >150 kDa [57]. They are however very large, and expensive to purchase and operate, rendering them currently somewhat impractical for general use.

2.2 Sample Preparation

1. Protein solution(s) to be analyzed.
2. Ammonium acetate >98 % pure.
3. 7.5 M ammonium acetate solution.
4. Formic acid.
5. Acetonitrile, HPLC grade.
6. Deionized water.
7. Amicon Ultra-0.5 centrifugal concentrators (Millipore, Billerica, USA).
8. Micro Bio-Spin P-6 chromatography columns (BioRad, Hemel Hempstead, UK).
9. Microcentrifuge.
10. Water bath or heating block.
11. UV–Vis spectrophotometer.
12. Timer/stopwatch.

2.3 Native Mass Spectrometry

1. Mass spectrometer (*see* Subheading 2.1) with associated services, installed as per the manufacturer's specifications.
2. IM-MS calibrant proteins [58] (Sigma Aldrich, Gillingham, UK).
3. Sodium Iodide 99.99 % (Sigma Aldrich, Gillingham, UK).
4. Caesium Iodide, analytical standard (Fluka).
5. AA stainless steel Dumont tweezers (Agar Scientific, Stansted, UK).
6. 2a stainless steel Dumont tweezers (Agar Scientific, Stansted, UK).
7. Column scribe (ceramic FSOT cutter) (Sigma Aldrich, Gillingham, UK).
8. Stereomicroscope.
9. 0.5–20 µL Eppendorf geLoader tips (Fisher Scientific).
10. Nanoelectrospray needles (New Objective). Alternatively make your own using materials 11–13 below [59].
11. Glass capillaries, 1.0 mm OD × 0.78 mm ID, either no filament or filament (Harvard Apparatus, Kent, UK).
12. Flaming/Brown P-97 micropipette puller (Sutter Instrument Co., Novato, CA).
13. Polaron range model SC7680 sputter coater (Quorum Technologies, Newhaven, UK).

3 Methods

3.1 Mass Spectrometer Modification for High-Molecular-Weight Transmission

1. Depending on the instrument (*see* Subheading 2.1), modifications may be needed to render it suitable for measuring high-mass protein complexes (*see* Subheading 2.1).
2. The major consideration is adequate focussing of the large ions such that they traverse through the vacuum stages of the mass spectrometer [14]. Improving the focussing of these ions is often required, depending on the mass spectrometer at hand, and is most readily achieved by increasing the pressure in the source ion guide regions at the inlet region of the instrument, by approximately an order of magnitude. This can be achieved through the introduction of a low flow of gas, reducing vacuum pumping speed, or the addition of a sleeve around the inlet [53–55].
3. For tandem-MS experiments where ions at high m/z need to be selected with a quadrupole analyzer, the radio-frequency generator for the quadrupole needs to be replaced with a unit that operates at lower frequency [53–55].

3.2 Sample Preparation for Native Mass Spectrometry

1. Native MS requires the protein complex of interest to be in a volatile buffer (*see* Note 1), at a concentration typically between 10 and 50 μM , in terms of monomer.
2. To exchange the sample into the buffer of choice, a range of devices can be employed that variously implement the techniques of dialysis, gel filtration, and ultrafiltration. The choice of technique depends on sample amount, concentration, and stability (*see* Note 2).
3. Determine the protein concentration of each sample destined for mass analysis by measuring absorbance at 280 nm, or by other means.
4. If the concentration of the sample is within 10–100 μM (monomer), and the sample volume $>20\ \mu\text{l}$, use Biospin P-6 columns for buffer exchange method.
5. Biospin columns are effectively a small gel filtration column. First equilibrate by washing the resin with the buffer of choice, and then load between 20 and 75 μl of the protein sample (following the manufacturer's instructions). The sample elutes from the column in the final spin whilst buffer components $<6,000\ \text{Da}$ are retained on the column (*see* Note 3).
6. Dilute samples can be concentrated and buffer-exchanged simultaneously by using centrifugal filtration devices, following the manufacturer's instructions.
7. As an example, for Amicon Ultra-0.5 load $<500\ \mu\text{l}$ of the sample into the spin filter device, after checking membrane integrity (*see* Note 4) and washing (*see* Note 5).

8. Spin the sample for 5 min and check flux (*see Note 6*). If concentration to 50 μl takes longer than 5 min, every few minutes gently mix the solution, by pipetting up and down. This can help prevent the protein concentration near the membrane reaching levels where precipitation could occur, thereby occluding the pores of the concentrator.
9. After concentrating to approximately 50 μl , top up to 500 μl with buffer, and recommence the concentration procedure. Repeat this process 3–5 times, to effectively dilute out the original buffer, stopping on the last cycle when the desired sample volume is attained.
10. Refer to alternative detailed protocols for further advice [59–64].

3.3 Sample Preparation to Examine Denatured Protein Complexes

1. Various methods exist for denaturing protein complexes to obtain accurate mass measurements of the constituent subunits [59, 62, 63].
2. The simplest approach is to first buffer exchange the sample into water using a Biospin column (*see Note 7*), and secondly add acetonitrile (50 % v/v) and formic acid (0.1 % v/v) (*see Note 8*).

3.4 Mass Calibration

1. Obtaining mass spectra of a calibrant is important to enable accurate mass measurement of the unknown, but also to test whether the mass spectrometer is operating to specification.
2. Dissolve caesium iodide (CsI) or sodium iodide (NaI) in water (acetonitrile 50 % v/v can be added), producing a solution <100 mg/ml, depending on the mass range of interest (*see Note 9*).
3. Briefly acquire data for the chosen salt solution (*see Note 10*, and refer to Subheading 3.5 if employing nanoESI). Ensure that peaks are observed over an m/z range spanning that expected for the protein complex to be subsequently analyzed (*see Note 11*).
4. The mass spectrum obtained can be used to directly calibrate the instrument prior to analysis, following the instrument manufacturer's instructions. Alternatively the calibration spectrum can be acquired independently of subsequent measurements and used to calibrate individual MS data files off-line (*see Note 12*).
5. After CsI/NaI infusion, remove and clean the sample cone by rinsing with water, and then sonicating for 20 min in solution of methanol:water:formic acid (45:45:10 v/v). For instruments without a readily removable sample cone, clean the exterior surface with a swab.

3.5 Nanoelectrospray and Transmission of Large Ions

1. Prepare a conductive borosilicate glass nanoelectrospray needle (*see Note 13*) for sample loading by mounting it in the holder using tweezers. Cut the “blunt” end of the needle using the

ceramic tile to obtain an appropriate length for the needle holder.

2. Trim the tip of the needle under a microscope using fine-tipped tweezers to obtain an orifice of approximately 10 μm diameter (*see Note 14*) [59].
3. Transfer 2 μl of buffer-exchanged and desalted sample into the prepared needle, using a gel-loading pipette tip. Flick the needle to transfer the sample towards the tip, and load the needle holder into the mass spectrometer.
4. Slowly apply a pressure of backing gas until a small droplet (<1 mm diameter) of sample is extruded from the needle (*see Note 15*).
5. Engage the needle holder such that the mass spectrometer source voltages are switched on, and the droplet is propelled into the instrument. If this does not occur, ensure that all safety interlocks are engaged, and voltages are on (*see Note 16*).
6. Adjust the needle position such that it is between 3 and 10 mm from the sample cone (*see Note 17* and Fig. 1).
7. Optimize the signal for stability, intensity, and narrowness of the peaks observed by altering the needle position, capillary

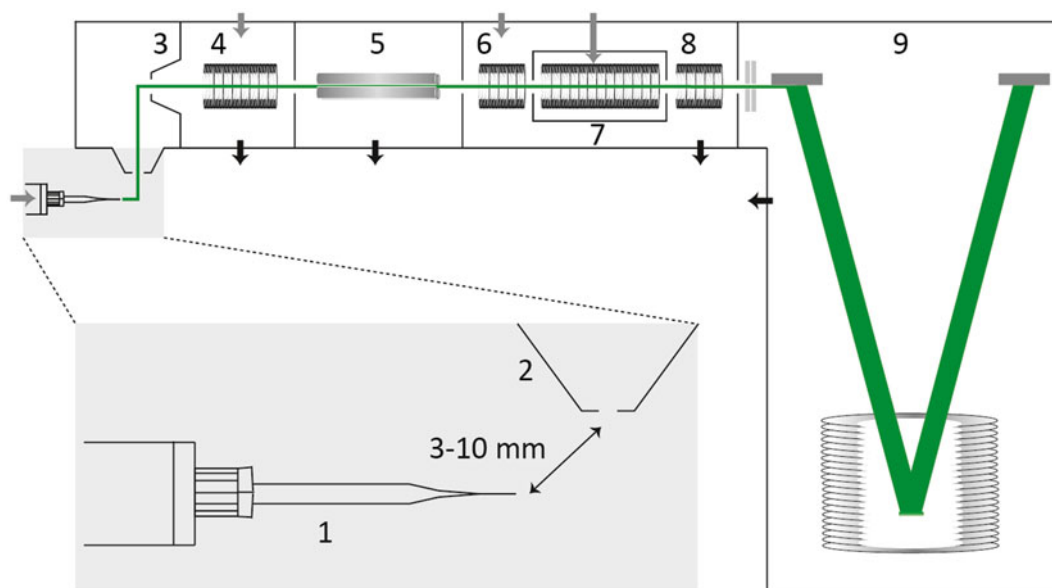


Fig. 1 Schematic of a Synapt-type Q-IM-ToF mass spectrometer. Sample is introduced into the vacuum of the instrument from the nanoESI needle **1** through the sample cone **2**. Ions (green) pass through the skimmer cone **3** and source ion guide **4** before reaching the quadrupole **5**. Here they can be selected according to m/z ratio, before travelling towards the trap collision cell **6**, IM cell **7**, and transfer collision cell **8**, before finally being separated in the ToF analyzer **9**. The *black arrows* indicate pumping, whereas the *grey arrows* denote where gas may be introduced. *Inset*: The needle is positioned at a distance of a few millimeters diagonal to the sample cone

voltage (1–2 kV), and backing gas pressure (*see Note 18*). Optimal conditions will vary for different protein samples, but in general higher backing gas pressures require higher voltages for stable electrospray. Backing pressures as low as possible are desirable, due to the slower sample consumption rate and smaller droplet sizes formed.

8. If the spray is stable and peaks are observed, further optimize settings by changing the sample and skimmer cone voltages, and source ion guide pressure (*see Note 19*).
9. The aim of these optimizations is to achieve the narrowest peaks possible and maximum ion current, and avoid unwanted dissociation (*see Note 20*). It is often not possible to achieve all of these at the same time, and acquiring a set of spectra exploring these conditions can be beneficial for downstream data analyses.
10. Refer to alternative detailed protocols for further advice [59–64].

3.6 Collision-Induced Dissociation

1. CID is generally carried out in-source or within a dedicated collision cell by increasing sample and/or skimmer cone voltages, or the acceleration voltage into the collision cell, respectively.
2. In either case, once a stable spray has been achieved (*see Subheading 3.5*), increase the appropriate voltage in relatively small increments (e.g., 10 V) whilst maintaining all other conditions (*see Note 21*). Ensure that the acquisition m/z range is wide enough to collect all dissociation products (*see Note 22*).
3. A voltage series such as this will enable the determination of the CID pathway and provide a means for assessing protein stability in the gas phase.

3.7 Tandem Mass Spectrometry (MS/MS) on a Q-ToF Instrument

1. Tandem MS is a refinement of the CID experiment (*see Subheading 3.6*), in which only selected ions are isolated for dissociation.
2. Identify a charge state from the complex of interest for further interrogation, change the mass spectrometer acquisition mode to MS/MS, and input the chosen m/z value (*see Note 23*). Acquire data with minimal acceleration into the collision cell to ensure that the correct m/z has been selected by the quadrupole (*see Note 24*).
3. Collecting data over an appropriate m/z range (*see Note 22*), increase the collision energy incrementally until CID occurs (*see Fig. 2*). Optimize the collision gas pressure for maximum intensity of fragment peaks and their resolution (*see Note 25*).
4. If insufficient CID occurs even at the highest accelerating voltage, two different experiments can be tried. A heavier collision gas (e.g., Xe, SF₆, *see Note 26*) can be used in the collision cell,

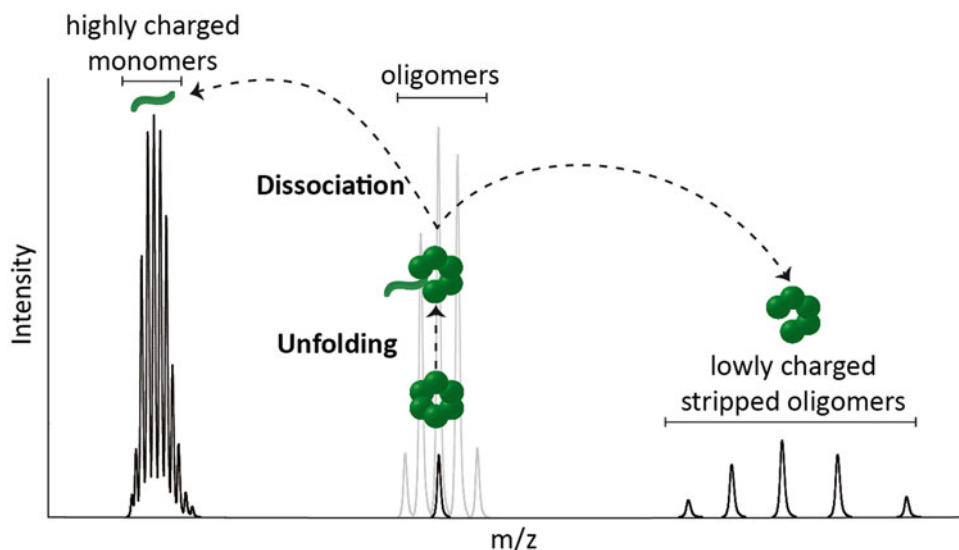


Fig. 2 Schematic tandem MS of a protein assembly. The native mass spectrum (*grey*) displays a series of charge states, one of which can be selected downstream interrogation by means of tandem MS. Upon collisional activation a subunit in the oligomer starts to unfold and is subsequently released, carrying away a large proportion of the charge from the complex. These expelled subunits therefore appear at low m/z values, with corresponding stripped oligomers at high m/z . Notably the separation between adjacent charge states is much greater for the stripped oligomers than for the parent oligomers, a property which can be exploited for deconvoluting complex spectra [27]

improving CID as more kinetic energy will be converted to internal energy per collision [14]. Alternatively, supercharging reagents yield higher charge states [60], which have higher dissociation efficiency during CID.

3.8 Ion Mobility Mass Spectrometry on a Synapt Instrument

1. IM provides an orthogonal dimension of separation, based on CCS, to the native MS experiment. For every feature observable in the m/z dimension of IM-MS data, an associated arrival time can be extracted (*see* Fig. 3).
2. In order to obtain CCS values representative of the native state of the protein (i.e., without deformation in the gas phase) it is important to minimize acceleration voltages prior to the IM cell.
3. Therefore, once a stable spray is achieved (*see* Subheading 3.5), lower the sample and skimmer cone voltages, IM bias voltage, and trap collision voltage while monitoring the arrival time, until no effect on the peak position that would reflect either a compaction or unfolding transition is observed (*see* Note 27) [25].
4. Once optimum voltages are found, obtain a series of spectra at a range of individual IM travelling-wave heights and/or velocities [65, 66].

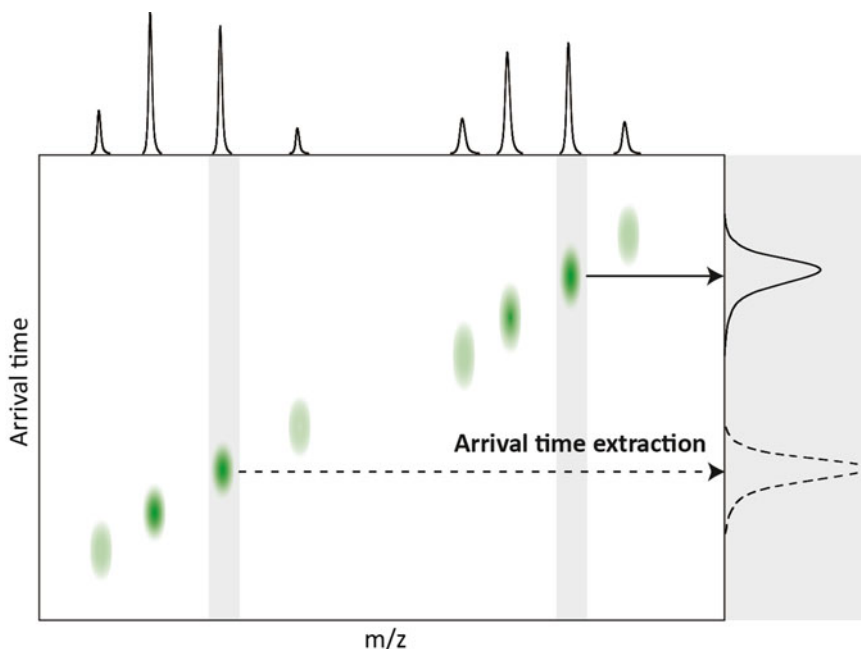


Fig. 3 Schematic IM-MS spectra. The combination of IM with native MS results in three-dimensional data, in which every feature resolved in m/z has an associated arrival time distribution. From this data the CCS of the analyte may be determined, typically after calibration with appropriate standards. This information is particularly useful for heterogeneous proteins, as illustrated here, where multiple forms are populated at equilibrium and for which traditional techniques would typically provide an ensemble average

5. Perform similar experiments on a selection of suitable calibrants (*see* **Note 28**). These should reflect the protein of interest, bracketing its estimated mass and mobility. A range of appropriate calibrants for native MS have been well characterized and are commercially available [58].
6. Calibrate the data by comparing the arrival times measured for the protein of interest and the selected calibrants, using detailed protocols [58, 66–68].
7. If monitoring collision-induced unfolding is desirable (*see* **Note 29**), increase either the sample or skimmer cone voltages, or trap collision voltage, by small increments (e.g., 10 V) and record the IM-MS data (*see* **Note 30**).

3.9 Elucidating Quaternary Dynamics by Monitoring Subunit Exchange

1. Monitoring subunit exchange requires that there is an observable mass difference between the two reacting proteins, such that they and anticipated hetero-oligomers can be identified. In the case of related proteins there may already be an appropriate mass difference; alternatively one can be introduced by the incorporation of a tag, or through metabolic labeling (*see* **Note 31**).

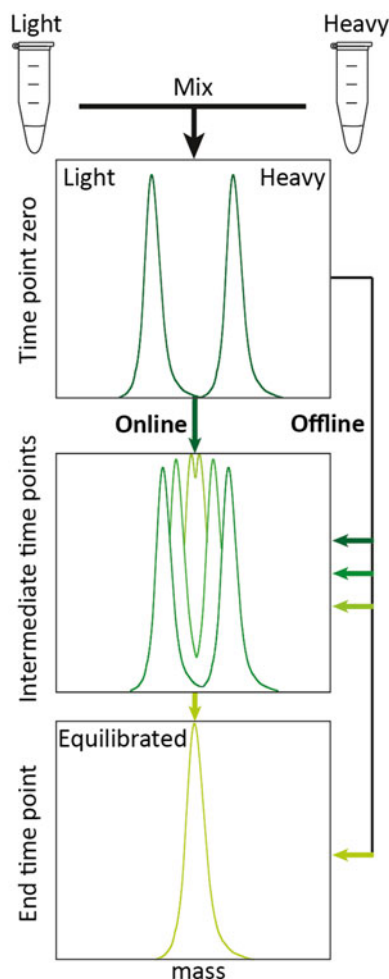


Fig. 4 Schematic of subunit exchange monitoring by means of native MS. Aliquots of two proteins of different mass are mixed and incubated. A spectrum (shown for clarity on a mass scale rather than m/z) as close as possible to time zero reveals the two masses, that of each homo-oligomer. The reaction mixture can then be followed either online in the nanoESI needle, or off-line. The mass spectra show the formation and evolution of hetero-oligomers, until an equilibrium distribution is reached and subunit exchange is no longer observable

2. Subunit exchange experiments can be carried out either off-line, where samples are incubated for different time points and then introduced individually into the mass spectrometer, or online, where a sample is sprayed over an extended period of time in a single needle and exchange monitored (*see* **Note 32** and Fig. 4).
3. Obtain native mass spectra for both the labeled and unlabeled proteins, identifying optimum instrument conditions. An overlay of these spectra will generate a hypothetical time zero for the subunit exchange data.

4. Mix equal amounts of unlabeled and labeled protein to make a reaction mixture (*see Note 33*), giving a final volume defined by the intended number of infusions/time points (approximately 5 μl /infusion), and a concentration that gives good-quality native MS data for the individual proteins. Start a timer.
5. As quickly as possible after mixing, remove 2 μl of sample and obtain a native MS spectrum (*see Note 34*), recording the time at which the acquisition was started.
6. For the online approach simply continue to spray the sample and acquire data for as long as possible, following the evolution of the mixture in the needle during its infusion. For long acquisition times it is advisable to load 5 μl of sample into the needle.
7. For the offline approach, incubate the remaining sample, removing 2 μl of the mixture and obtaining native mass spectra at the desired time points (*see Note 35*), until the sample is used up or exchange is complete, and no difference is observed between successive time points (*see Note 36*).
8. The rate of exchange for the protein of interest or the desire to examine the temperature dependence of quaternary dynamics often demands performing subunit exchange experiments at elevated or reduced temperatures.
9. Online experiments can be conducted at temperatures 10–100 °C by using a nanoESI needle holder equipped with temperature regulation [69].
10. For off-line experiments, the sample can be incubated in a water bath or temperature-controlled block, with aliquots removed and infused, assuming that the exchange occurring during data acquisition is negligible (*see Note 37*).
11. The observation that for many proteins subunit exchange is very slow at low temperature enables cooling to be used as a means for “quenching” the reaction. This can be exploited by removing 5 μl aliquots from the incubating mixture and transferring them into individual prechilled tubes at planned time points, and keeping them on ice prior to MS analysis (*see Note 38*).

4 Notes

1. Salts and nonvolatile buffer components (e.g., phosphate, NaCl, Tris, HEPES) need to be removed from native MS samples as their presence often results in a loss of signal intensity and resolution. Ammonium acetate (e.g., 200 mM) is suitable for many proteins, though alternatives such as ammonium bicarbonate can be used [60]. The ionic strength required to

maintain stability in solution, and to achieve an optimal mass spectrum, is protein dependent, so screening a range of buffer concentrations is beneficial (between 10 mM and 1 M). If the protein of interest requires specific solution additives in order to remain stable, it is preferable to reduce their concentration as far as possible to ensure minimal adduct formation. Membrane proteins have stringent solubility requirements, but can be examined both in detergent [70] and detergent-free environments [71, 72].

2. For small amounts of sample, which is typical of many native MS investigations, the buffer exchange devices of choice are typically Biospin P-6 columns or low-volume centrifugal filtration devices such as Amicon Ultra-0.5. The most effective means for buffer exchange is gel filtration by means of FPLC, but is only practical for sample amounts of >0.5 mg.
3. The majority of samples only require a single run through a Biospin column; however repeating the procedure (using new Biospin columns each time) is beneficial in some cases.
4. To check for tears in the membrane load with 500 μ L of buffer and spin for 5 min. The membrane is intact if some buffer is retained in the upper reservoir.
5. Load device with 500 μ L of buffer and spin for 5 min. This should remove traces of packing stabilizers. Repeat twice. Once washed, either use immediately or leave buffer in the device to prevent the membrane from drying out.
6. The flux will mainly be dependent on sample concentration, centrifugal force, temperature, sample particulates, and the viscosity of the solution. To increase flux where applicable, lower the protein concentration, increase centrifugal force, filter the sample prior to concentration, and increase the temperature.
7. The presence of ammonium acetate or other buffer reduces the impact of the formic acid and hence the likelihood of obtaining a good-quality denatured mass spectrum.
8. The addition of acetonitrile and formic acid disrupts non-covalent interactions, thereby dissociating the complex and unfolding the monomeric subunits in the solution phase.
9. CsI and NaI form large clusters $[(\text{CsI})_n\text{Cs}]^+$ and $[(\text{NaI})_n\text{Na}]^+$ (in positive ion mode), where n is an integer, that can span a wide m/z range. For high-molecular-weight complexes it is often advisable to use CsI for calibration, as its clusters extend to higher mass ranges. NaI is more appropriate for lower mass ranges due to the smaller spacing between successive peaks.
10. CsI and NaI should provide high ion currents; therefore <60 s of data is generally sufficient. Excessive infusion will lead to

fouling of the inlet optics, causing subsequent decreased sensitivity or carryover.

11. Vary the source ion guide pressure, and sample and extractor cone voltages to optimize transmission. In Q-ToF instruments, also optimize accelerating voltages into the collision cell, and gas pressure within.
12. Cs, Na, and I are mono-isotopic and therefore each cluster gives rise to only a single peak. Many peaks can however be present within the spectrum including doubly and triply charged ions. Some cluster ions are more abundant than others, due to preferred geometrical arrangements [73].
13. Alternatively an automated infusion device can be used, such as the Nanomate (Advion Biosciences) [74], in which case skip to **step 5**. The parameters on this robotic platform require optimization from one protein to the next, which can be time and sample consuming, but once established is very useful for the screening of, for example, ligand binding [46], or solution conditions [75].
14. Make sure that the cut creates a straight edge. The size of the orifice is something that can be optimized based on the solution viscosity. In general, smaller orifice sizes are preferable due to the lower sample consumption rates, and the smaller initial droplet sizes leading to less adduction.
15. If no droplet appears, first verify that the gas supply to the instrument is on. If so, trim the needle slightly, or try another needle. If the problem persists, the sample may be too concentrated.
16. In general positive-ion mode provides better signal-to-noise ratios; however negative-ion mode is also suitable for native MS studies [76].
17. The position of the needle relative to the cone has significant effects on the quality of the mass spectra due to its influence on the electric field gradient, and droplet evaporation time. If the needle is moved too close to the cone it can cause the conductive coating at the needle tip to be stripped off and loss of signal.
18. If a stable spray cannot be obtained prepare a fresh needle, as nanospray needles are prone to variations in shape, and heterogeneous coating. If the needle clogs repeatedly, or frequent adjustments of backing gas pressure are required, dilute the sample, and repeat.
19. Increasing the sample and skimmer cone voltages, and acceleration voltage into the collision cell, can help resolve large mass peaks by collisional “cleaning” [25]. However, if the voltages are too high, unwanted dissociation can occur.

Higher pressures can improve collisional focussing, and therefore increase signal. However, there is an optimum: excessive pressure can cause scattering resulting in broader peaks, and ultimately signal drop-off.

20. If unwanted dissociation is unavoidable, consider varying the buffer composition to promote solution-phase stability (which may be the root of the problem), and explore gas-phase charge-reduction and evaporative-cooling approaches [77, 78].
21. It is practical to first scout the voltage range, trying the high and low extremes, to see at what point (if at all) dissociation is observed, before acquiring a larger data set at many voltages.
22. A broader mass range acquisition is often required compared to the native spectrum. This is because the dissociated subunits will appear at very low m/z (comparatively small and highly charged) and stripped oligomers will be at high m/z (loss of a disproportionate amount of charge in comparison to mass).
23. If the quadrupole calibration is inaccurate, then adjust the inputted m/z value until the correct peak is selected. Similarly, if more than one charge state has been selected by the quadrupole the selection window is too wide. Optimize the quadrupole resolution voltages until only the m/z peak of interest is selected.
24. Selection of the precursor ion occurs in the quadrupole. This is positioned upstream of the collision cell (*see* Fig. 1) wherein activation can cause collisional cleaning, and hence cause the peak to shift to lower m/z prior to ToF analysis. Care therefore needs to be taken to ensure that the appropriate m/z window for selection by the quadrupole is inputted into the software, such that the intended peak (rather than an adjacent one) is selected.
25. If no stripped oligomers are observed, try increasing the gas pressure in the collision cell. These species tend to be more likely to drift off-axis and the increased pressure will help to radially confine them on a trajectory leading to the detector.
26. These gases are relatively expensive, so are not ideal for everyday use.
27. These minimally activating conditions will result in broader peaks in the m/z dimension, due to incomplete removal of solvent and buffer ions. In other words, to get arrival time data that reflects the solution-phase structure, the quality of the mass spectrum will inevitably suffer.
28. It is vital that the gas pressure in the IM cell and voltages applied to lenses between the IM cell and the detector are identical and maintained throughout all measurements between the calibrants and the protein of interest, as even minor changes will result in inaccurate CCS determination.

29. It is desirable to perform this collision-induced unfolding experiment by tandem MS because otherwise effects such as ligand and charge loss can act to confound the data and produce ambiguous results [43].
30. The sample and skimmer cones, in addition to the trap collision cell, are all located upstream of the IMS cell (*see* Fig. 1). Inducing protein unfolding in these regions of the instrument therefore causes a measurable increase in the CCS. Acceleration into the transfer collision cell, which is located after the IM cell, will not cause an observable difference in CCS, as activation in this region occurs after IM separation.
31. Metabolic labeling can be achieved by expression in the presence of either, or both, ^{15}N or ^{13}C . Before this is carried out it is advisable to calculate expected peak positions when these labels are incorporated, based on experimentally determined charge-state distributions for the protein. This can be readily achieved by determining the number of carbons or nitrogens in the protein, adjusting the average molecular mass accordingly, and calculating the m/z value for the relevant charge states. The labeling scheme selected based on these simulations should provide maximal spacing between equivalent charge states for anticipated hetero-oligomers while avoiding overlap between adjacent charge states. If the use of either isotope appears feasible, consider cost: ^{15}N -ammonium chloride is significantly cheaper than ^{13}C -glucose.
32. The online approach has the advantage of providing many data points, and doing so in real time. However, it requires a stable spray over the entire duration of the exchange experiment, making it impractical if completion is not reached in <1 h.
33. It may be desirable to choose a ratio other than 1:1, to skew the distribution of hetero-oligomers formed, thereby reducing the complexity of the spectra [47].
34. Pipetting and concentration measurement errors mean that the “heavy” and “light” proteins will typically not be present at equal intensities in the mass spectrum. If the discrepancy is small this can be readily accounted for when the data is analyzed. However if it is large, it may be preferable to make a new reaction mixture taking into account the intensity information from the mass spectra.
35. An alternative to manual nanoESI infusion is to employ the Nanomate (Advion), having programmed a method in which the sample is infused automatically at regular time points [79].
36. The aim is to collect data across the complete exchange time course. However, the largest changes in the mass spectra will occur early in the experiment; therefore it is useful to distribute the time points accordingly. When the approximate timescale

exchange is unknown, it is advisable to carry out an exploratory experiment with a small amount of sample, taking time points as dictated by the evolving data set, to provide an initial estimate of exchange rate.

37. Check that this assumption is fair by comparing data obtained at the beginning and end of the acquisition.
38. If subunit exchange is fast (<10 min) it is best to pipette 5 μ l aliquots of the ice-cold mixture into separate tubes and incubate them all at the exchange temperature. At each time point a separate tube can then be removed and placed on ice (or even snap-frozen). In this way subunit exchange reactions that are complete in 60 s have been successfully monitored [80].

References

1. Shendure J, Aiden EL (2012) The expanding scope of DNA sequencing. *Nat Biotechnol* 30:1084–1094
2. Mann M, Kulak NA, Nagaraj N, Cox J (2013) The coming age of complete, accurate, and ubiquitous proteomes. *Mol Cell* 49:583–590
3. Benesch JLP, Ruotolo BT (2011) Mass spectrometry: come of age for structural and dynamical biology. *Curr Opin Struct Biol* 21:641–649
4. Konermann L, Vahidi S, Sowole MA (2013) Mass spectrometry methods for studying structure and dynamics of biological macromolecules. *Anal Chem* 86:213–232
5. Heck AJR (2008) Native mass spectrometry: a bridge between interactomics and structural biology. *Nat Methods* 5:927–933
6. Hyung S-J, Ruotolo BT (2012) Integrating mass spectrometry of intact protein complexes into structural proteomics. *Proteomics* 12:1547–1564
7. Konijnenberg A, Butterer A, Sobott F (2013) Native ion mobility-mass spectrometry and related methods in structural biology. *Biochim Biophys Acta* 1834:1239–1256
8. Schmidt C, Robinson CV (2014) Dynamic protein ligand interactions - insights from MS. *FEBS J* 281(8):1950–1964
9. Sharon M (2013) Structural MS pulls its weight. *Science* 340:1059–1060
10. Beveridge R, Chappuis Q, Macphree C, Barran P (2013) Mass spectrometry methods for intrinsically disordered proteins. *Analyst* 138:32–42
11. Barrera NP, Robinson CV (2011) Advances in the mass spectrometry of membrane proteins: from individual proteins to intact complexes. *Annu Rev Biochem* 80:247–271
12. Sali A, Glaeser R, Earnest T, Baumeister W (2003) From words to literature in structural proteomics. *Nature* 422:216–225
13. Winston RL, Fitzgerald MC (1997) Mass spectrometry as a readout of protein structure and function. *Mass Spectrom Rev* 16:165–179
14. Benesch JLP, Ruotolo BT, Simmons DA, Robinson CV (2007) Protein complexes in the gas phase: technology for structural genomics and proteomics. *Chem Rev* 107:544–3567
15. Hilton GR, Benesch JLP (2012) Two decades of studying non-covalent biomolecular assemblies by means of electrospray ionization mass spectrometry. *J R Soc Interface* 9:801–816
16. Marcoux J, Robinson CV (2013) Twenty years of gas phase structural biology. *Structure* 21:1541–1550
17. Loo JA (1997) Studying noncovalent protein complexes by electrospray ionization mass spectrometry. *Mass Spectrom Rev* 16:1–23
18. Snijder J, Rose RJ, Veesler D et al (2013) Studying 18 MDa virus assemblies with native mass spectrometry. *Angew Chem Int Ed* 52:4020–4023
19. Alberts B (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* 92:291–294
20. Ruotolo BT, Robinson CV (2006) Aspects of native proteins are retained in vacuum. *Curr Opin Chem Biol* 10:402–408
21. Breuker K, McLafferty FW (2008) Stepwise evolution of protein native structure with electrospray into the gas phase, 10(-12) to 10(2) S. *Proc Natl Acad Sci U S A* 105:8145–18152
22. Fenn J, Mann M, Meng C et al (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* 246:64–71

23. Konermann L, Ahadi E, Rodriguez AD, Vahidi S (2013) Unraveling the mechanism of electrospray ionization. *Anal Chem* 85:2–9
24. Wilm M, Mann M (1996) Analytical properties of the nanoelectrospray ion source. *Anal Chem* 68:1–8
25. Benesch JLP (2009) Collisional activation of protein complexes: picking up the pieces. *J Am Soc Mass Spectrom* 20:341–348
26. Shukla AK, Futrell JH (2000) Tandem mass spectrometry: dissociation of ions by collisional activation. *J Mass Spectrom* 35:1069–1090
27. Benesch JLP, Aquilina JA, Ruotolo BT et al (2006) Tandem mass spectrometry reveals the quaternary organization of macromolecular assemblies. *Chem Biol* 13:597–605
28. Jurchen JC, Williams ER (2003) Origin of asymmetric charge partitioning in the dissociation of gas-phase protein homodimers. *J Am Chem Soc* 125:2817–2826
29. Taverner T, Hernández H, Sharon M et al (2008) Subunit architecture of intact protein complexes from mass spectrometry and homology modeling. *Acc Chem Res* 41:617–627
30. Aquilina JA, Benesch JLP, Bateman OA et al (2003) Polydispersity of a mammalian chaperone: mass spectrometry reveals the population of oligomers in alpha B-crystallin. *Proc Natl Acad Sci U S A* 100:10611–10616
31. Hall Z, Politis A, Robinson CV (2012) Structural modeling of heteromeric protein complexes from disassembly pathways and ion mobility-mass spectrometry. *Structure* 20:1596–1609
32. Wysocki VH, Jones CM, Galhena AS, Blackwell AE (2008) Surface-induced dissociation shows potential to be more informative than collision-induced dissociation for structural studies of large systems. *J Am Soc Mass Spectrom* 19:903–913
33. Uetrecht C, Rose RJ, van Duijn E (2010) Ion mobility mass spectrometry of proteins and protein assemblies. *Chem Soc Rev* 39:1633–1655
34. Lanucara F, Holman SW, Gray CJ, Evers CE (2014) The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics. *Nat Chem* 6:281–294
35. McLean JA (2009) The mass-mobility correlation redux: the conformational landscape of anhydrous biomolecules. *J Am Soc Mass Spectrom* 20:1775–1781
36. Politis A, Stengel F, Hall Z et al (2014) A mass spectrometry-based hybrid method for structural modeling of protein complexes. *Nat Methods* 11:403–406
37. Thalassinou K, Pandurangan AP, Xu M et al (2013) Conformational states of macromolecular assemblies explored by integrative structure calculation. *Structure* 21:1500–1508
38. Zhong Y, Hyung SJ, Ruotolo BT (2012) Ion mobility-mass spectrometry for structural proteomics. *Expert Rev Proteomics* 9:47–58
39. Giles K, Pringle SD, Worthington KR et al (2004) Applications of a travelling wave-based radio-frequency only stacked ring ion guide. *Rapid Commun Mass Spectrom* 18:2401–2414
40. McCammon MG, Scott DJ, Keetch CA (2002) Screening transthyretin amyloid fibril inhibitors: characterization of novel multiprotein, multiligand complexes by mass spectrometry. *Structure* 10:851–863
41. Hopper JT, Oldham NJ (2009) Collision induced unfolding of protein ions in the gas phase studied by ion mobility-mass spectrometry: the effect of ligand binding on conformational stability. *J Am Soc Mass Spectrom* 20:1851–1858
42. Hyung SJ, Robinson CV, Ruotolo BT (2009) Gas-phase unfolding and disassembly reveals stability differences in ligand-bound multiprotein complexes. *Chem Biol* 16:382–390
43. Rabuck JN, Hyung SJ, Ko KS et al (2013) Activation state-selective kinase inhibitor assay based on ion mobility-mass spectrometry. *Anal Chem* 85:6995–7002
44. Sharon M, Robinson CV (2007) The role of mass spectrometry in structure elucidation of dynamic protein complexes. *Annu Rev Biochem* 76:167–193
45. Chevreux G, Atmanene C, Lopez P, Ouazzani J, Van Dorsselaer A, Badet B, Badet-Denisot MA, Sanglier-Cianferani S (2011) Monitoring the dynamics of monomer exchange using electrospray mass spectrometry: the case of the dimeric glucosamine-6-phosphate synthase. *J Am Soc Mass Spectrom* 22:431–439
46. Keetch CA, Bromley EHC, McCammon MG et al (2005) L55P Transthyretin accelerates subunit exchange and leads to rapid formation of hybrid tetramers. *J Biol Chem* 280:41667–41674
47. Sobott F, Benesch JLP, Vierling E, Robinson CV (2002) Subunit exchange of multimeric protein complexes: real-time monitoring of subunit exchange between small heat shock proteins by using electrospray mass spectrometry. *J Biol Chem* 277:38921–38929
48. Hilton GR, Hochberg GKA, Laganowsky A et al (2013) C-terminal interactions mediate the quaternary dynamics of alpha B-crystallin. *Philos Trans R Soc Lond B Biol Sci* 368(1617):20110405

49. Kozlovski VI, Donald LJ, Collado VM et al (2011) A TOF mass spectrometer for the study of noncovalent complexes. *Int J Mass Spectrom* 308:118–125
50. Tahallah N, Pinkse M, Maier CS, Heck AJR (2001) The effect of the source pressure on the abundance of ions of noncovalent protein assemblies in an electrospray ionization orthogonal time-of-flight instrument. *Rapid Commun Mass Spectrom* 15:596–601
51. Rose RJ, Damoc E, Denisov E, Makarov A, Heck AJR (2012) High-sensitivity Orbitrap mass analysis of intact macromolecular assemblies. *Nat Methods* 9:1084–1086
52. Lössl P, Snijder J, Heck AJR (2014) Boundaries of mass resolution in native mass spectrometry. *J Am Soc Mass Spectrom* 25(6):906–917
53. Sobott F, Hernandez H, McCammon MG (2002) A tandem mass spectrometer for improved transmission and analysis of large macromolecular assemblies. *Anal Chem* 74:1402–1407
54. van den Heuvel RHH, van Duijn E, Mazon H et al (2006) Improving the performance of a quadrupole time-of-flight instrument for macromolecular mass spectrometry. *Anal Chem* 78:7473–7483
55. Chernushevich IV, Thomson BA (2004) Collisional cooling of large ions in electrospray mass spectrometry. *Anal Chem* 76:1754–1760
56. Pringle SD, Giles K, Wildgoose JL et al (2007) An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument. *Int J Mass Spectrom* 261:1–12
57. Li H, Wolff JJ, Van Orden SL, Loo JA (2013) Native top-down electrospray ionization-mass spectrometry of 158 kDa protein complex by high-resolution Fourier transform ion cyclotron resonance mass spectrometry. *Anal Chem* 86:317–320
58. Bush MF, Hall Z, Giles K et al (2010) Collision cross sections of proteins and their complexes: a calibration framework and database for gas-phase structural biology. *Anal Chem* 82:9557–9565
59. Hernandez H, Robinson CV (2007) Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nat Protoc* 2:715–726
60. Lorenzen K, van Duijn E (2001) Native mass spectrometry as a tool in structural biology. In: *Current protocols in protein science*. Wiley, New York
61. Kirshenbaum N, Michalevski I, Sharon M (2010) Analyzing large protein complexes by structural mass spectrometry. *J Vis Exp* e1954
62. Sanglier S, Atmanene C, Chevreux G, Dorselaer AV (2008) Nondenaturing mass spectrometry to study noncovalent protein/protein and protein/ligand complexes: technical aspects and application to the determination of binding stoichiometries. *Methods Mol Biol* 484:217–243
63. Yin S, Loo JA (2009) Mass spectrometry detection and characterisation of noncovalent protein complexes. *Methods Mol Biol* 492:273–282
64. Campuzano I, Giles K (2011) Nanospray ion mobility mass spectrometry of selected high mass species. *Methods Mol Biol* 790:57–70
65. Zhong Y, Hyung SJ, Ruotolo BT (2011) Characterizing the resolution and accuracy of a second-generation traveling-wave ion mobility separator for biomolecular ions. *Analyst* 136:3534–3541
66. Ruotolo BT, Benesch JL, Sandercock AM et al (2008) Ion mobility-mass spectrometry analysis of large protein complexes. *Nat Protoc* 3:1139–1152
67. Smith DP, Knapman TW, Campuzano I et al (2009) Deciphering drift time measurements from travelling wave ion mobility spectrometry-mass spectrometry studies. *Eur J Mass Spectrom* (Chichester, Eng) 15:113–130
68. Thalassinos K, Grabenauer M, Slade SE et al (2009) Characterization of phosphorylated peptides using traveling wave-based and drift cell ion mobility mass spectrometry. *Anal Chem* 81:248–254
69. Benesch JLP, Sobott F, Robinson CV (2003) Thermal dissociation of multimeric protein complexes by using nanoelectrospray mass spectrometry. *Anal Chem* 75:2208–2214
70. Laganowsky A, Reading E, Hopper JT, Robinson CV (2013) Mass spectrometry of intact membrane protein complexes. *Nat Protoc* 8:639–651
71. Hopper JT, Yu YT, Li D et al (2013) Detergent-free mass spectrometry of membrane protein complexes. *Nat Methods* 10:1206–1208
72. Leney AC, McMorran LM, Radford SE, Ashcroft AE (2012) Amphipathic polymers enable the study of functional membrane proteins in the gas phase. *Anal Chem* 84:9841–9847
73. Campana JE, Colton RJ, Wyatt JR et al (1984) Ultra-high mass spectrometry. *Appl Spectrosc* 38:430–432
74. Van Pelt CK, Zhang S, Henion JD (2002) Characterization of a fully automated nanoelectrospray system with mass spectrometric detection for proteomic analyses. *J Biomol Tech* 13:72–84

75. Zhong Y, Feng J, Ruotolo BT (2013) Robotically assisted titration coupled to ion mobility-mass spectrometry reveals the interface structures and analysis parameters critical for multiprotein topology mapping. *Anal Chem* 85:11360–11368
76. Allen SJ, Schwartz AM, Bush MF (2013) Effects of polarity on the structures and charge states of native-like proteins and protein complexes in the gas phase. *Anal Chem* 85:12055–12061
77. Hopper JT, Sokratous K, Oldham NJ (2012) Charge state and adduct reduction in electrospray ionization-mass spectrometry using solvent vapor exposure. *Anal Biochem* 421: 788–790
78. Bagal D, Kitova EN, Liu L et al (2009) Gas phase stabilization of noncovalent protein complexes formed by electrospray ionization. *Anal Chem* 81:7801–7806
79. Painter AJ, Jaya N, Basha E et al (2008) Real-time monitoring of protein complexes reveals their quaternary organization and dynamics. *Chem Biol* 15:246–253
80. Baldwin AJ, Lioe H, Robinson CV et al (2011) alphaB-crystallin polydispersity is a consequence of unbiased quaternary dynamics. *J Mol Biol* 413:297–309

INDEX

A

Automation 64, 142, 150, 151, 153, 212,
217, 235, 250, 292–298, 315

B

Baculovirus 91–113, 115, 116, 118, 197–208
Beamlines 142, 150, 212, 224,
225, 227, 228, 234–246, 250, 251, 256–259, 266,
267, 273, 292–294, 332
Beer–Lambert relationship 285
Bioreactors 130, 161, 176, 182, 192,
193, 199, 202, 205, 208

C

CCS. *See* Collisional cross section (CCS)
CD. *See* Circular dichroism (CD)
Cell-free 129–139, 190
 expression 171–194, 335
Chemical shift
 anisotropy 285, 287, 332, 342, 343
 assignment 303–315, 318, 321
 shift value 305, 306, 312, 317, 318, 321, 322
Chinese hamster ovary (CHO) 130, 132, 135–137
Cholesteryl hemisuccinate (CHS) 163, 169, 204
Circular dichroism (CD) 38, 255, 281
 spectroscopy 38, 255–275
Co-infection 92, 95, 102–105, 112
Collisional cross section (CCS) 353, 360, 361,
366, 367
Co-translational 37, 39, 52, 55,
92, 115, 129, 130, 138, 143, 172, 182, 184–188,
349, 352
Coupled transcription-translation 135
Cre-mediated recombination 106, 107
Cre recombinase 64, 66, 69, 76, 105, 107–109, 113
Critical assessment of protein structure prediction 44
Crystal dehydration 223, 244, 251
Crystallization 5, 6, 40, 48, 141–153,
159, 197, 211–228, 234–238, 240, 242–244, 250,
251, 256, 259–267, 273
Crystallography 5, 10, 37, 38, 45,
47, 141, 142, 150, 197, 211, 233, 236, 242, 255,
256, 279

D

Databases 4, 7, 9–14, 16, 22, 24,
26, 37–39, 46, 48, 49, 52–54, 246, 342
DDM. *See* *n*-Dodecyl β -D-maltoside (DDM)
Detection of protein 102, 135, 181, 184, 208, 333
Detergents 48, 160, 169,
172, 177, 178, 182, 184–186, 189–192, 198,
203–208, 211–215, 217–219, 225–228, 236,
242, 256, 258, 268, 332, 335, 364
Differential scanning fluorimetry (DSF) 144,
145, 151
Diffraction data 152, 246
 collection 235, 236, 244, 250
Dipolar couplings 332–334, 338, 343, 344
Disordered proteins 7, 12, 35–55, 279, 287
n-Dodecyl β -D-maltoside (DDM) 163, 169,
174, 178, 184, 185, 198, 199, 203, 204, 207,
215, 217
DSF. *See* Differential scanning fluorimetry (DSF)
Dynamic light scattering (DLS) 143–145, 236

E

Electronic laboratory notebook (ELN) 24, 25, 32
Electrospray ionization (ESI) 350
Escherichia coli 52, 63–88, 92, 94, 98,
99, 106–108, 115, 130, 132, 134, 160, 168, 170,
173, 175, 177–179, 181, 199, 206, 217
ESI. *See* Electrospray ionization (ESI)
Expression screening 197–208

F

Far-UV region 25, 256, 265–267, 272
Fluorescence-detection size exclusion chromatography
 (FSEC) 160, 162–163,
166–167, 170, 198, 199, 202–207
Force field 305

G

Gas-phase 350–354, 359, 360, 366
GenBank 3, 5, 131
Gene ontology 6, 8
GFP. *See* Green fluorescent protein (GFP)

Goniometry239
G-protein coupled receptors (GPCRs)13,
14, 48, 159–178, 181, 184–191, 197,
198, 202, 222
Green fluorescent protein (GFP)160, 161,
163–170, 173, 179, 181, 182, 186, 187, 190, 193,
197–199, 208
Guinier law.....295

H

HEK 293 cells. *See* Human embryonic kidney (HEK) 293 cells
HEK293S-GnTI.....116, 117, 121–126
HEK293T.....116, 117, 119, 121–123
High-throughput.....3, 16, 40, 47, 52,
53, 64, 144, 204, 205, 208, 213, 215–224,
233–251, 256, 258, 277–298, 349–368
Homologous recombination.....94, 98, 103,
105, 106, 164, 199
HSQC experiment309, 311
Human embryonic kidney (HEK) 293 cells.....116, 198

I

IDP. *See* Intrinsically disorder/unstructured protein (IDP)
Immobilized metal affinity chromatography
(IMAC)109, 173, 189, 206
In-gel fluorescence.....198, 202–205
Insect cells 91–113, 116, 119, 129, 161, 167, 168, 197–208
In situ crystal dehydration.....223
In situ diffraction.....233–251
Intrinsically disorder/unstructured protein
(IDP)7, 12, 35–55
Ion channels48, 197, 331, 335
Ion mobility spectrometry.....352–353, 360–361

K

K562 cells130, 132
Kifunensine117, 123, 125
Kratky plot.....285–287

L

Laboratory information management system
(LIMS)24, 32, 33
LCP. *See* Lipidic cubic phase (LCP)
Ligation independent cloning102
LIMS. *See* Laboratory information management system
(LIMS)
Lipidic cubic phase (LCP)212, 213,
219–222, 228, 239

M

Machine-learning methods43, 44
Magic angle spinning (MAS).....332, 333,
336–338, 342–344
MALLS. *See* Multi-Angle Laser Light Scattering (MALLS)

Mass spectrometry.....268, 349–368
Mass spectrum.....352, 357, 360, 364, 366, 367
Matrix seeding.....152
Membrane proteins12, 48,
84, 115, 129, 152, 160, 171, 176, 197–208,
211–228, 234, 258, 261–264, 272–273,
331–345
Microbatch142, 237
Microfocus.....225, 228, 235, 250
Molecular envelope.....279
Monte Carlo methods307
Multi-angle laser light scattering
(MALLS)212, 214, 215, 217
Multi-gene64, 66
Multiple sequence alignment.....53, 102
Multi-protein complex91–113

N

Nanodisc (ND).....176, 177, 186,
188–191, 236, 335, 343
assembly.....189–190
Native mass spectrometry.....349–368
ND. *See* Nanodisc
Near-UV region256, 257,
268, 269, 272
Network-anchoring.....321–324
N-linked glycosylation.....116
NMR. *See* Nuclear magnetic resonance (NMR)
NOE. *See* Nuclear Overhauser effect (NOE)
NOESY.....305, 306, 308, 309,
315, 316, 318–323, 325, 326
Nuclear magnetic resonance (NMR).....5, 7, 9,
10, 12, 37–39, 42, 47–49, 52, 53, 255, 256, 272,
279, 303–326, 331–345
spectrometer304, 336
Nuclear Overhauser effect (NOE)303, 305,
315, 316, 318–325

P

PDB. *See* Protein Data Bank (PDB)
PDSD. *See* Proton driven spin diffusion (PDSD)
PEG. *See* Polyethylene glycol (PEG)
PEI. *See* Polyethyleneimine (PEI)
Plate screening.....223–225, 228
Polyethylene glycol (PEG)142, 161,
165, 217, 228, 259
Polyethyleneimine (PEI)117, 121, 124, 125
Post-translational modifications37, 52, 55,
92, 115, 143, 349, 352
Protein
crystals51, 141, 142,
144–147, 150–153, 211–213, 217–219, 221–225,
228, 234–237, 239, 242–244, 246, 247, 250, 251,
256, 331
secondary structure prediction54, 311

structure.....3–16, 40, 48, 54, 211,
234, 255, 279, 303–306, 313, 314, 319, 320, 325,
331, 335, 338, 350
Protein Data Bank (PDB).....3–6, 37, 141, 149, 255
Proteoliposomes188, 335
Proteomicelles171
Proton driven spin diffusion (PDSD).....334,
338–341, 344
Purification tag.....64, 70, 119, 121, 181, 190

Q

Quality control (QC).....109, 192, 256, 268, 273

R

Radiation damage.....235, 244, 294
Resonance.....309, 310, 320,
321, 333, 334, 342, 343
assignment.....306–309, 314–316,
319, 321, 338, 340, 341, 344

S

Saccharomyces cerevisiae.....159–170
SAXS. *See* Small angle x-ray scattering (SAXS)
SBKB. *See* Structural Biology Knowledgebase (SBKB)
SCOP. *See* Structural Classification of Proteins (SCOP)
Screening.....53, 85, 95, 96, 98, 102,
103, 112, 116, 142–148, 151, 159–161, 173,
182, 184, 186, 188, 190, 191, 193, 198, 202,
203, 205–207, 223–225, 227, 228, 235–237, 239,
240, 242, 244, 250, 256, 257, 261–264, 273, 350,
364, 365
Size-exclusion chromatography.....190, 198,
207, 212–215
Small angle x-ray scattering (SAXS)10, 49, 277–298
Small scale expression test.....103–105, 119–121
Solid-state NMR.....316, 318, 332–338, 342, 344, 345
Solution NMR316, 318, 332–334
Specific labelling of amino acids.....334

Spodoptera frugiperda.....130, 198
Stable cell lines119–123
Structural Biology Knowledgebase
(SBKB)13–15, 17
Structural Classification of Proteins
(SCOP).....10–11, 38
Structural genomics.....14, 16, 40, 47–49,
145–147, 151, 234
Structural proteomics21–34, 349–368
Structure
annotation.....3–17
modelling.....282, 289–293, 298
Suspension culture.....93, 97, 101, 104, 120, 123–125
Synchrotron.....24, 212, 219, 223–225,
227, 234–236, 243, 277–298

T

Target selection.....24, 34, 40, 47
Thermal shift assay.....146
Time-of-flight (ToF).....354, 358, 366
Torsion angle dynamics315, 326
Transient transfection.....119, 120, 123–125
Transmembrane.....14, 48, 51, 136, 169, 334, 343

U

Unfolded protein49, 145, 287, 352

V

Vapor diffusion method.....142, 213, 217–219
Virus amplification.....100–102

X

X-ray
crystallography.....5, 10, 37, 38, 45,
47, 141, 197, 211, 255, 256, 279
diffraction142, 152, 211, 234–238, 279
structure.....5, 37, 38

